



Interactive Evidence Detection

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr. rer. nat.

vorgelegt von
Chris Stahlhut, M.Sc.
geboren in Buchholz in der Nordheide

Tag der Einreichung: 05.06.2020

Tag der Disputation: 15.07.2020

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt
Prof. Dr. Johannes Fürnkranz, Linz
Prof. Dr. Jens Ivo Engels, Darmstadt

Darmstadt 2021
D17

Stahlhut, Chris: Interactive Evidence Detection
Darmstadt, Technische Universität Darmstadt
URN: urn:nbn:de:tuda-tuprints-191546
Jahr der Veröffentlichung der Dissertation auf TUpriints: 2021
Tag der mündlichen Prüfung: 15.07.2020

Veröffentlicht unter CC BY 4.0 International
<https://creativecommons.org/licenses/>

Wissenschaftlicher Werdegang des Verfassers¹

06/06–09/10	Bachelor of Science (B.Sc.) in Informatik an der Universität Hamburg
10/10–03/15	Master of Science (M.Sc.) in Informatik an der Universität Hamburg
12/15–09/16	Wissenschaftlicher Mitarbeiter im Fachgebiet Ubiquitous Knowledge Processing (UKP-Lab) der Technischen Universität Darmstadt
10/16–09/19	Doktorand am Graduiertenkolleg KRITIS und am Fachgebiet Ubiquitous Knowledge Processing (UKP-Lab) der Technischen Universität Darmstadt

¹Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt

Abstract

Without evidence, research would be nearly impossible. Whereas a conjecture in mathematics must be proven logically until it is accepted as theorem, most research disciplines depend on empirical evidence. In the natural sciences, researchers conduct experiments to create the most objective evidence to evaluate hypotheses. In the humanities and social sciences, most evidence is extracted from textual sources. These can be news articles spanning several decades or transcribed interviews. However, not every document or interview contains statements that support or contradict a hypothesis, causing a time-intensive search that might be sped up with modern natural language processing techniques.

Finding evidence —or evidence detection— is a fast growing field that is currently gaining relevance because of the increased focus on detecting fake news. Some research focusses not only on evidence detection, but also on linking evidence to hypotheses, or evidence linking. Other work aims at speeding up the decision processes regarding whether a hypothesis is valid or not. Yet another focus of research in evidence detection aims at finding evidence in medical abstracts. Although these approaches are promising, their applicability to research in the humanities and social sciences has not yet been evaluated. Most evidence detection and evidence linking models are also static in nature. Usually, we first create a large dataset in which text snippets are labelled as evidence. This dataset is then used to train and evaluate different models that do not change after their initial training. Furthermore, most work assumes that all users interpret evidence in a similar way so that a single evidence detection or evidence linking model can be used by all users.

This PhD project aims at evaluating whether modern natural language processing techniques can be used to support researchers in the humanities and social sciences in finding evidence so that they can evaluate their hypotheses. We first investigated how real users search for evidence and link this to self-defined hypotheses. We found that there is no canonical user; some users define hypotheses first and then search for evidence. Others search for evidence first and then define hypotheses. We also found that the interpretation of evidence varies between different users. Similar hypotheses are supported by different pieces of evidence, and the same evidence can be used to support different hypotheses. This means that any evidence detection model must be specific to a single user.

User-specific evidence detection models require a large amount of data, which is labour-intensive to create. Therefore, we investigate how much data is necessary until an interactively trained evidence detection model outperforms a well generalising, state-of-the-art model. In our evaluation, we found that an evidence detection model, which had first been trained on external data and then been fine-tuned interactively, requires only a few training documents to yield better results than a state-of-the-art model trained only on the external data.

Regarding the practical benefit of this research, we built an annotation or coding tool allowing users to label sentences as evidence and link these pieces of evidence to self-defined hypotheses. We evaluated this tool, named EDoHa (Evidence Detection fOr Hypothesis vAlidation), in a user study with a group of students and one with colleagues from the research training group KRITIS. EDoHa and the data to pre-train evidence detection and evidence linking models are published under an open source licence so that researchers outside the research training group can also benefit from it. This project contributes not only to evidence detection and natural language processing, but also to research methodologies in qualitative text-based research.

Zusammenfassung

Ohne Evidenzen ist Wissenschaft nahezu unmöglich. Während eine Vermutung in der Mathematik logisch bewiesen werden muss, ehe sie als Theorem akzeptiert wird, werden in den meisten Wissenschaften empirische Evidenzen benötigt. In den Naturwissenschaften werden Experimente durchgeführt, die dazu dienen, die möglichst objektivsten Evidenzen zu finden, um damit Hypothesen zu überprüfen. In den Geistes- und Sozialwissenschaften hingegen werden die Evidenzen häufig aus Textquellen erhoben. Dies können Zeitungsartikel aus mehreren Jahrzehnten oder Jahrhunderten sein, oder aber auch transkribierte Interviews. Jedoch enthält nicht jedes Dokument oder jedes Interview auch Aussagen, die eine Hypothese be- oder widerlegen, wodurch die Suche danach sehr zeitaufwändig ist und sich womöglich mit Techniken des modernen Natural Language Processing beschleunigen lässt.

Das Finden von Evidenzen —oder Evidence Detection— ist ein schnell wachsendes Feld, das derzeit stark an Bedeutung gewinnt, da Zeitungsenten oder Fake News stärker in den Fokus gelangen. Oftmals wird neben Evidence Detection auch das Verknüpfen der Evidenzen mit Hypothesen, oder Evidence Linking betrachtet. Es finden sich Arbeiten, die darauf abzielen, Entscheidungen schneller treffen zu können. Ein weiteres Anwendungsgebiet von Evidence Detection ist das Finden von Evidenzen in medizinischen Abstracts. Auch wenn diese Ansätze bereits vielversprechend sind, wurde ihre Anwendbarkeit im Kontext der Geistes- und Sozialwissenschaften bisher nicht evaluiert. Zudem sind die Evidence Detection und Evidence Linking Modelle statisch. Für gewöhnlich wird zunächst ein großer Datensatz erstellt, in dem Textstellen als Evidenz oder nicht Evidenz markiert sind. Dieser Datensatz wird dann genutzt, um unterschiedliche Modelle zu trainieren und zu evaluieren, wobei die Modelle jeweils nur einmal trainiert werden und sich danach nicht mehr ändern. Ebenso gehen die Arbeiten davon aus, dass alle User Evidenzen gleich interpretieren und somit ein einziges Evidence Detection oder Evidence Linking Modell von allen Usern genutzt werden kann.

Ziel dieses Promotionsprojekts ist, zu evaluieren, ob moderne NLP-Techniken genutzt werden können, um Wissenschaftlerinnen und Wissenschaftler in den Geistes- und Sozialwissenschaften dabei zu helfen, Evidenzen zu finden und damit ihre Hypothesen zu evaluieren. Wir erforschten zunächst, wie echte User Evidenzen finden und mit selbst definierten Hypothesen verknüpfen. Hierbei stellten wir fest, dass es keinen kanonischen User gibt. Manche Personen definieren zunächst Hypothesen und suchen dann nach Evidenzen. Andere hingegen sammeln zunächst Evidenzen und definieren die Hypothesen später. Wir konnten ebenfalls feststellen, dass die Interpretation von Evidenzen variiert. Ähnliche Hypothesen werden mit unterschiedlichen Evidenzen unterstützt und dieselben Evidenzen werden genutzt, um unterschiedliche Hypothesen zu unterstützen. Daraus folgt, dass ein Evidence Detection Modell nutzerspezifisch sein muss.

Nutterspezifische Evidence Detection Modelle benötigen eine große Menge an Daten, welche aufwändig zu erstellen sind. Daher stellt sich die Frage, ob man nicht ein auf anderen Daten vortrainiertes Evidence Detection Modell nutzen kann, welches bereits einen kleinen Nutzen bietet und damit das Cold-Start Problem adressiert. Zudem ist offen, wie viele Trainingsdaten notwendig sind, bis ein interaktiv trainiertes Evidence Detection Modell besser ist als ein gut generalisierendes State-of-the-Art Evidence Detection Modell. In unserer Evaluation stellten wir fest, dass ein Evidence Detection Modell, welches zunächst auf externen Daten vortrainiert und dann interaktiv verfeinert wurde, bereits nach sehr wenigen Datenpunkten bessere Ergebnisse lieferte als ein State-of-the-Art Evidence Detection Modell, welches nur auf den externen Daten trainiert wurde.

Bezüglich des praktischen Nutzens, haben wir ein Annotations- oder Codingwerkzeug

entwickelt, das von einzelnen Usern lernt, welche Sätze Evidenzen sind und welche Hypothesen diese unterstützen. Der Nutzen dieses Werkzeugs, genannt EDoHa (Evidence Detection fOr Hypothesis vAlidation), wurde bereits in einer Studie mit Studierenden der Geistes- und Sozialwissenschaften validiert und wurde in Zusammenarbeit mit Kolleginnen und Kollegen tiefer gehend evaluiert. EDoHa, ebenso wie Daten um Evidence Detection und Evidence Linking Modelle vortrainieren zu können, ist inzwischen unter einer Open Source Lizenz veröffentlicht, was es auch Nicht-Mitgliedern des Graduiertenkollegs ermöglicht, von diesem Promotionsprojekt zu profitieren. Dieses Projekt liefert somit nicht nur einen Beitrag zur Evidence Detection und Natural Language Processing, sondern bietet auch einen Beitrag zu der Forschungsmethodik für qualitative Textarbeit.

Acknowledgments

This thesis has been supported by the German Research Foundation (DFG) as part of the Research Training Group KRITIS No. GRK 2222/1.

First and foremost, I would like to thank Prof. Dr. Iryna Gurevych for giving me the opportunity to pursue a PhD in the Ubiquitous Knowledge Processing Lab. Second, I would like to thank Prof. Dr. Jens Ivo Engels and Kathrin Reichert in particular for their support in finding participants for the user studies which enabled my research. Furthermore, I would like to extend my sincere gratitude to Prof. Dr. Johannes Fürnkranz, Prof. Dr. Karsten Weihe, Prof. Dr. Carsten Binnig, and Prof. Sebastian Faust, Ph.D. for reviewing my PhD thesis and their questions during my defense. I would also like to thank Christian Stab and Johannes Daxenberger for their guidance and numerous rounds of feedback for everything I wrote; without them, I would certainly not be writing these lines.

I would also like to thank my colleagues from UKP for their fruitful discussions, input to my work, and contributions in doing my research and building EDoHa. Specifically, I would like to thank Richard Eckart de Castilho, Nina Kolmar, Ute Winchenbach, Benjamin Schiller, Ji-Ung Lee, and Andreas Hanselowski for reading my drafts and supporting me in developing EDoHa.

KRITIS was more than a group of PhD students doing research in an interdisciplinary context, it developed into a group of friends who shared their experience in undergoing this project. I would not have been able to finish my research without their support and challenges to my approaches. Most notably, I would like to thank Arturo Crespo, Marcus Dombois, Alice Knauf, Nadja Thießen, Benedict Vianden, Marcel Siegler, Allegra Baumann, and Stefanie Eifert for their input, but especially for their friendship.

Contents

1	Introduction	1
1.1	Contributions	5
1.2	Publication Record	6
1.3	Thesis Organisation	8
2	Research Methodologies and Evidence	11
2.1	Evidence in Philosophy of Science	12
2.2	Evidence in Natural Language Processing	14
2.3	Evidence in History	18
2.4	Evidence in Sociology	20
2.5	Chapter Conclusion	21
3	Automatic Support for Evidence Detection and Hypothesis Validation	23
3.1	Current Research in Evidence Detection	24
3.2	Fact-Checking and Claim Verification	28
3.3	Argument Mining	33
3.4	Other Related Research	35
3.5	Chapter Conclusion	37
4	Human Strategies for Evidence Detection	39
4.1	User Study	40
4.2	EDoHa	42
4.3	User Behaviour	45
4.4	User Feedback	54
4.5	Data	55
4.6	Discussion	58
4.7	Chapter Conclusion	59
5	Machine Learning for Evidence Detection	61
5.1	Direct Training	62
5.2	Transfer Learning	68
5.3	Interactive Learning	80
5.4	Chapter Conclusion	84
6	Machine Learning for Evidence Linking	87
6.1	Data	88
6.2	Models	88
6.3	Hyper-Parameter Tuning	90
6.4	Results	92
6.5	Discussion	95

6.6	Chapter Conclusion	97
7	Interactive Evidence Detection	99
7.1	Interactive Evidence Detection with EDoHa	100
7.2	Pre-training Data and Models	103
7.3	Evaluation	108
7.4	Discussion	111
7.5	Chapter Conclusion	113
8	Conclusion	115
8.1	Lessons Learned	117
8.2	Speeding up Fact-Checking with Interactive Evidence Detection	118
8.3	Future Work	120
A	Consent Form and Task Description for User Study	137
A.1	Consent Form	138
A.2	Task Descriptions for User Studies and Feedback forms	140
B	Source Listing	143
C	Hyper-Parameter Optimisation for ML for EL	145

Chapter 1

Introduction

Without evidence most research would be impossible. Whereas a mathematical conjecture cannot be validated empirically but must be proven logically to be accepted as a theorem, a hypothesis in the sciences or other fields of research must be evaluated based on evidence. If the evidence does not support the hypothesis, or even contradicts it, then the hypothesis is most likely wrong. And for accepted hypotheses, it is the evidence that gives them credibility to the point that they are valid, meaning they should be accepted as being true.

Evidence is often seen as a “neutral arbiter” (Kelly, 2016) in which increasing amounts of evidence will lead to a convergence of opinion among competent readers. A competent reader needs to be able to understand and interpret the evidence, e.g. understand the meaning of the metrics used in comparing two systems as well as the validity of the experimental setup. This view is apparent in the court of law as well as in scientific research. The same applies to a large part of research in Natural Language Processing (NLP) and Machine Learning (ML). Although NLP as well as ML have sub-fields which aim at creating knowledge from logical deduction (e.g. Dayan, 1992; Sonoda and Murata, 2019) most contemporary research is empirical in nature.

Evidence is generally seen as a fact that is used in an argument to support or contradict a claim or hypothesis (Daston, 1991). This means, it is always part of a relation and does not stand alone; evidence must always be evidence of or for something else. Generally, this something refers to a hypothesis, such as “*the defendant has killed the victim*”, or a method “*topic information improves the performance of a particular argument mining method*”.

What constitutes evidence, as well as the approach to finding it, varies greatly depending on the field of research or the evidence’s purpose. In a court of law, evidence might be the fingerprints of a defendant on the knife used to kill the victim (Kelly, 2016). In NLP, to evaluate the aforementioned benefit of topic information, researchers would first build the method with topic information **a** and the method without topic information **b** and then compare **a** and **b** by evaluating their performance in carefully designed experiments. The purpose of experiments is to create the most informative, reliable, and objective evidence currently possible. The evidence as to whether the topic information is beneficial takes the form of the performance measurements of the methods **a** and **b** produced during the experiments.

Scholars of history spend vast amounts of time in archives reading through countless documents. Evidence to them can be found in figures which demonstrate a common scene in life, or statements made in public and private records. For instance, the statement “*The economy and, in particular, the energy-intensive companies, which are an important element of the value chain in Germany, also need competitive energy prices*” can be

used to support the hypothesis “*The nuclear phase-out is fundamentally connected to the industry*”¹. This hypothesis can then be used in a larger context of understanding the vulnerabilities and criticalities involved in the nuclear phase-out.

This thesis draws heavily from personal interactions at the research training group KRITIS. KRITIS consists of thirteen doctoral candidates with backgrounds varying from history, through sociology, political science and philosophy to civil engineering and informatics. Its research focusses on different aspects of critical infrastructure, such as the electrical grid or railway network, or the policies around it and the effects of natural catastrophes. From this context we will define our use cases which we will use to develop our research questions. Use case 1 shows a historian analysing the political discourse around the Chernobyl and the Fukushima catastrophes with the purpose of understanding the critical relationships of nuclear energy and the resulting vulnerabilities to the society.

Use Case 1. The historian starts their research by selecting archives to visit in search for relevant documents. They then visit these archives and collect potentially relevant documents, such as the minutes of plenary proceedings of the German parliament after the Chernobyl disaster and Fukushima catastrophe. While reading these minutes, the historian notices that there are many references to the industry and the cost involved in a nuclear phase-out. After marking a few such instances, the historian then formulates the hypothesis “*Corporate profit motive hampers the nuclear phase-out.*” They then connect this hypothesis with the previously found statements, which act as evidence in supporting the researcher’s hypothesis. Afterwards, the researcher continues reading the minutes and notices that there are several references to modifications of the nuclear power grid. This leads the researcher to formulate the hypothesis “*The nuclear power grid must be restructured for a nuclear phase-out.*” As before, the researcher links the previously found pieces of evidence to the hypothesis and continues reading.

In the social sciences, researchers often search for evidence in transcripts of interviews. They do so by using an *annotation* or *coding tool* which allows them to highlight snippets of text and organise them so that they can draw conclusions from it.² Use case 2 shows a sociologist who is researching the effect of cruise ship tourism on the infrastructure of small to mid-sized cities.

Use Case 2. The sociologist begins their research by defining one or two cities as case studies and then creates a list of candidates to interview. During a field trip, the sociologist conducts multiple interviews and then transcribes them. The sociologist then starts analysing the transcripts and marks statements, such as “*The city is tightened between the sea and the mountain.*” or “*We are planning to extend our public transport network.*”, for future use. They then link these statements to different hypotheses, such as “*Geographical and historical conditions play an important role.*” and “*The city government introduces measures to cope with the growing number of tourists.*” The sociologist then continues the analysis.

In these use cases, we can identify five tasks in which NLP can support the researcher, which figure 1.1 illustrates. Namely *document retrieval* to find relevant documents, *automated speech recognition* to transcribe interviews into text, *evidence detection* to find the

¹Many of the examples used in this thesis are taken from pieces of evidence labelled by users in analysing the political discourse of the German parliament around the topic of nuclear energy. We then translated the statements from German to English.

²In this thesis we will refer to text coding as annotation to keep the terminology consistent with the NLP community.

relevant pieces of evidence within the documents, *evidence linking* to link the pieces of evidence to hypotheses, and *claim verification* to determine whether or not the hypothesis is valid. Document retrieval is a mature research area within *information retrieval* (Manning, Raghavan, and Schütze, 2008) in which the user inputs a query and the machine returns a collection of documents ranked by relevance. Automated speech recognition is the task of converting human verbal utterances into text, for instance to transcribe text or to enable down-stream tasks, such as speech translation (Salesky, Sperber, and Black, 2019) or a verbal interface with applications. Extracting evidence from text, or Evidence Detection (ED) is an up-and-coming task within NLP with varying goals. Some research focusses on the medical context (Mayer, Cabrio, and Villata, 2018; Shardlow et al., 2018), while others focus on a debating context (Orbach et al., 2019). Evidence Linking (EL) is so far only being treated as a sub-step in ED, e.g. the context-dependent stage described by Rinott et al. (2015), or not at all (Shnarch et al., 2018). Reasoning about the validity of a hypothesis, or Hypothesis Validation (HV), is a sub-task of Claim Verification (CV), not in the sense that HV is a necessary step in CV, but in that although all hypotheses are claims, not all claims are hypotheses. The goal of CV in NLP is to be able to tell a user whether a controversial claim is valid or not (Thorne et al., 2018a; Thorne et al., 2018b). As in our example, the machine would first collect the relevant documents, then find the evidence that would either support or contradict the claim, and finally make a decision whether the evidence is strong enough to conclude that the claim is valid or not. In this thesis, we will refer to CV, when addressing the general task and related literature and to HV when addressing claims that are hypotheses in research. For instance, the statement “*Barack Obama was not born in the United States*” is a general claim, whereas the aforementioned “*Corporate profit motive hampers the nuclear phase-out*” is a research hypothesis. The differences lie in the context in which they are formulated and the ease of access to supporting or contradicting evidence. Evidence regarding the first claim can be found with relative ease in public documents, whereas finding evidence regarding the hypothesis requires careful reading of statements with knowledge about the speaker’s background and possible motives.

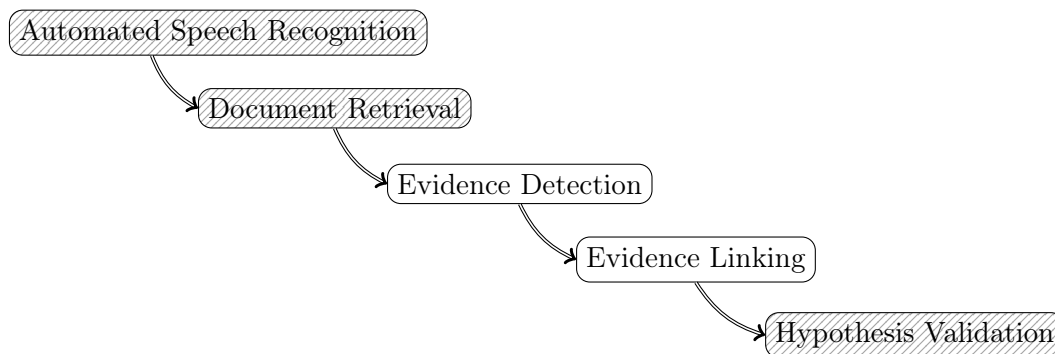


Figure 1.1: The five identified NLP tasks necessary to automatically validate hypotheses in a pipeline model. The tasks with the gray striped background are not considered in this thesis.

We decided not to investigate document retrieval and automated speech recognition, because the former is already a very mature area of research and the latter is only relevant to one of our two use cases; although sociologists oftentimes transcribe interviews, historians rarely rely on spoken sources. Using CV to reason about the validity of a hypothesis is not feasible; in research, evidence regarding a hypothesis is rarely stated directly in a source. Instead, it requires knowledge about the context, the reliability of a potential source, and social conditions. Therefore, validating hypotheses still requires the interpretation and

insight of a human researcher. In historical research in particular, the notion of evidence is more subjective and relies more on interpretation than researchers in the natural sciences are used to (Lloyd, 2008). Furthermore, there are multiple different schools of thought which form a framework of interpretations, such as *Historical materialism* and *Subaltern studies*. We therefore focus on supporting researchers in validating their hypotheses not by aiming at solving the HV task, but by supporting them in finding evidence. This means, we support them in the sub-ordinate tasks of ED and EL so that they can easier and faster decide whether or not a hypothesis is supported or contradicted by a sufficient amount of evidence to be considered valid or invalid. However, the subjectivity and differences between the schools of thought, in particular, do not allow for ED and EL to be treated as a general and user-independent task. This means, each user has their own personalised model which should be trained interactively, because researchers in history and the social sciences are not experts in NLP and ML. Based on these challenges and especially the requirement for both ED and EL models to be trained interactively, we developed our research questions.

RQ 1: How do researchers in the humanities and social sciences validate their hypotheses? To support researchers, we must first understand how they work, because otherwise any support without consideration of their approach to work might miss the point. For example, do scholars formulate their hypotheses first and then search for evidence, or do they first look for relevant sources and then develop their hypotheses based on them?

RQ 2: How well do machine learning-based methods work for ED? There are many different ways in which Machine Learning (ML) can be used to support a user. For instance, we can use existing data to train a large state-of-the-art model for ED, we can let the user first annotate a large amount of data before using an ED model, or we can let the user train an ED model interactively.

RQ 3: How well do machine learning-based methods work for EL? Similarly to the previous question, there are many different approaches one could follow in supporting researchers in linking evidence to hypotheses. We can let an individual user interactively train an EL model or use large out-of-domain data to train a user-independent model.

RQ 4: How do researchers benefit from interactively trained ED and EL models in their research? Simulated users and metrics have a limited quality of predicting what does and doesn't work when presented to real users. A high quality in a quantitative evaluation does not necessarily translate into a large benefit for users; conversely, even if a method does not perform well in quantitative evaluations, it can still be beneficial. Hence, to investigate the actual benefit of such methods, we need to integrate interactively trainable ED and EL models into a tool and evaluate the benefit with real users.

1.1 Contributions

This thesis makes contributions not only to NLP and ED but also offers contributions to the methodology of historical research as well as social and political sciences.

Natural Language Processing and Evidence Detection

❶ There is no agreement on evidence across users.

In NLP, we often assume that a task has a gold-standard solution, meaning that it can be solved globally. This is reinforced by having multiple people annotate the same data and then extract the solution with the highest agreement. However, when it comes to evidence, especially in an interactive scenario, this is not the case. The evaluation of agreement between different users on similar or identical hypotheses demonstrates that users disagree with one another. This means that in an interactive setting, any ED or EL model must adapt to the individual user.

❷ We provide a more realistic evaluation setting for interactive ED than random down-sampling and the datasets used in our experiments.

Many experiments attempt to evaluate how much training data is necessary before a model’s performance degrades drastically by randomly down-sampling the training data (e.g. Schulz et al., 2019). However, this does not take into account the bias introduced from the source documents. A single document often only covers a limited amount of aspects and topics; for instance, a report on the dangers of a nuclear fallout most likely does not contain anything related to the problems of nuclear waste management. If data from multiple documents is then down-sampled randomly, all aspects and topics might still be present. This can lead to an overestimation of the quality of a model trained on the reduced amount of training data and therefore of its behaviour when presented to actual users. We built more realistic simulations based on observations of the behaviour of actual users, which gives us a better understanding of how our methods work when presented to users.

❸ A simple model can outperform a state-of-the-art ED model when given in-domain training data.

The current trend in NLP is to build large pre-trained models, such as BERT (Devlin et al., 2018) which generalise better than the previous state-of-the-art. But when approaching new topics for which we have little training data, the quality of their generalisation is not comparable to a much simpler method that was trained on very little data that is topic-relevant.

❹ It often takes little in-domain data to outperform a state-of-the-art ED model trained on out-of-domain data.

In addition to our previous contribution, we also investigated how much training data is necessary to outperform a well generalising, state-of-the-art ED method. If a large amount of in-domain training data is required, then the labour involved in creating the training data might not be worth it. However, we found that in many cases and a large shift in domain, the training data from a single document is sufficient to outperform a state-of-the-art model trained on out-of-domain data.

Joint contributions to NLP, ED, and research methodology in History, Social, and Political Science

⑥ There is no canonical user for ED and HV.

Research can be conducted in many ways. Some argue that one must define the hypotheses first and then set out to find evidence to validate them. Others argue that a researcher must form their hypotheses from the evidence at hand (Andersen and Hepburn, 2016). However, the practice of conducting research most often falls somewhere in-between. In this thesis, we deliver quantitative evidence that the majority of users do not fully follow either of these approaches. After an initial definition, the user revises the hypothesis, because it fits the evidence better. When reading a collection of documents, some users link each found piece of evidence immediately to a hypothesis, while others work in phases of finding evidence and linking the evidence to hypotheses.

⑥ An annotation tool that learns from the user which sentence is evidential and which hypothesis a piece of evidence is connected to.

EDoHa (Evidence Detection fOr Hypothesis VALidation) is an annotation tool which allows historians or other researchers to find sources, or a social or political scientist to do their text annotation. It allows the user to label sentences as evidence quickly and to link them to self-defined hypotheses. Furthermore, as the user labels more and more sentences and creates more links between hypotheses and evidence, EDoHa can then use this data to train internal and user-dependent ED and EL models. These models are then used to make suggestions of sentences that might be evidential and which hypotheses they might be linked to. We do this to speed up the labour-intensive process of collecting and curating evidence while improving its quality.

⑦ Software and data to pre-train ED and EL models in English as well as German.

To avoid the cold-start problem, we provide training data for both ED and EL that allows the user to benefit immediately from the suggestions of the system. Furthermore, we also machine translated the training data from English to German, with minimal loss in performance, to allow a wider variety of researchers to benefit from the results of this thesis.

1.2 Publication Record

Parts of this thesis have been published at peer reviewed conferences and workshops.

- Chris Stahlhut, Christian Stab, and Iryna Gurevych (Aug. 2018). “Pilot Experiments of Hypothesis Validation Through Evidence Detection for Historians”. In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. First Biennial Conference on Design of Experimental Search & Information Retrieval Systems. Vol. 2167. CEUR Workshop Proceedings. Bertinoro, Italy, pp. 83–89. URL: <http://ceur-ws.org/Vol-2167/paper7.pdf>
- Chris Stahlhut (Aug. 2018). “Searching Arguments in German with ArgumenText”. In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. First Biennial Conference on Design of Experimental

Search and Information Retrieval Systems. Vol. 2167. CEUR Workshop Proceedings. Bertinoro, Italy, p. 104. URL: <http://ceur-ws.org/Vol-2167/short7.pdf>

- Chris Stahlhut (Nov. 2019b). “Interactive Evidence Detection: Train State-of-the-Art Model out-of-Domain or Simple Model Interactively?” In: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Hong Kong, China: Association for Computational Linguistics, pp. 79–89. DOI: 10.18653/v1/D19-6613. URL: <https://www.aclweb.org/anthology/D19-6613>
- Chris Stahlhut (Oct. 2019a). “Combatting Disinformation via Interactive Evidence Detection”. In: *Proceedings of the First Conference on Truth and Trust Online*. London, UK, p. 9. URL: https://truthandtrustonline.files.wordpress.com/2019/09/paper_9.pdf

We published additional work at peer-reviewed conferences and in peer-reviewed journals, as well as contributions to chapters and conference reports. These publications are outside the scope of this thesis.

- Chris Stahlhut, Nicolás Navarro-Guerrero, Cornelius Weber, and Stefan Wermter (2016). “Interaction in Reinforcement Learning Reduces the Need for Finely Tuned Hyperparameters in Complex Tasks”. In: *Kognitive Systeme 3*. DOI: <http://dx.doi.org/10.17185/dupublico/40718>
- Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych (June 2018a). “ArgumenText: Searching for Arguments in Heterogeneous Sources”. In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, USA: Association for Computational Linguistics
- Jens Ivo Engels, Jochen Monstadt, Marcus Dombois, Sybille Frank, Chris Stahlhut, and Tina Enders (2019). “Urban Infrastructures: Criticality, Vulnerability and Protection. Report of the International Conference of the Research Training Group KRITIS at Technische Universität Darmstadt, Germany”. In: *Urban Infrastructures: Criticality, Vulnerability and Protection*. Ed. by Jens Ivo Engels. Darmstadt, Germany. URL: <http://tubiblio.ulb.tu-darmstadt.de/113656/>
- Jens Ivo Engels, Kristof Lukitsch, Marcel Müller, and Chris Stahlhut (Aug. 2018a). “Criticality”. In: *Key Concepts for Critical Infrastructure Research*. Springer VS, Wiesbaden, pp. 11–20. ISBN: 978-3-658-22919-1 978-3-658-22920-7. URL: <http://tubiblio.ulb.tu-darmstadt.de/106805/>
- Jens Ivo Engels, Kristof Lukitsch, Marcel Müller, Chris Stahlhut, Stephanie Eifert, Alice Knauf, Nadja Thiessen, Ivonne Elsner, Andreas Huck, Manas Marathe, Arturo Crespo, Marcus Dombois, and Jan Henning (Aug. 2018b). “Relations between the Concepts”. In: *Key Concepts for Critical Infrastructure Research*. Springer VS, Wiesbaden, pp. 45–52. ISBN: 978-3-658-22919-1 978-3-658-22920-7. URL: <http://tubiblio.ulb.tu-darmstadt.de/106805/>

1.3 Thesis Organisation

The remainder of this thesis is separated into six chapters, each addressing different aspects of this work. Figure 1.2 illustrates their content, dependencies, and how the chapters relate to our contributions.

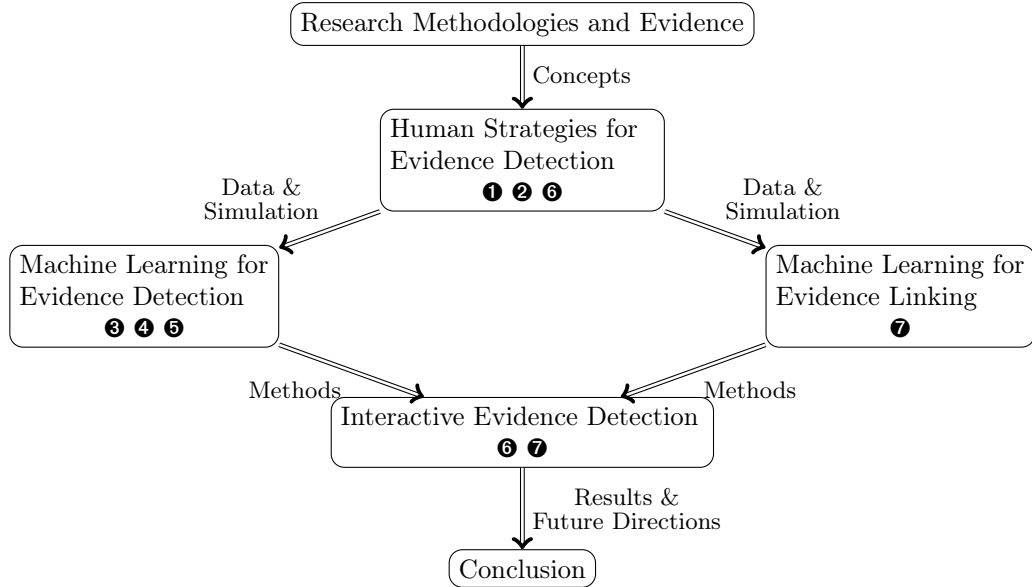


Figure 1.2: Structure of the remainder of this thesis and the dependencies between the chapters. The circled numbers show the relations to the contributions of this thesis.

Chapter 2 Research Methodologies and Evidence:

In this chapter, we describe the background of our work. This includes hypothesis validation in philosophy of science and what constitutes evidence in NLP and other research disciplines. We focus on the differences between NLP and our use case disciplines and, especially, the differences within each use case discipline.

Chapter 3 Automatic Support for Evidence Detection and Hypothesis Validation:

This chapter describes current approaches using automated support to find evidence or to automatically (in)validate user-defined claims, as well as other related research such as argument mining and interactive machine learning.

Chapter 4 Human Strategies for Evidence Detection:

We investigate in this chapter how humans develop hypotheses and validate them with evidence to answer our first research question (RQ 1). We show through two user studies with students from the humanities, social, and political sciences that there is no one-size-fits-all approach to ED. Furthermore, after comparing the evidence that users linked to similar and identical hypotheses, we conclude that support for ED must be user-specific. This chapter is based on Stahlhut, Stab, and Gurevych (2018) and contains contributions ① and ②, as well as parts of contribution ⑥.

Chapter 5 Machine Learning for Evidence Detection:

We address our second research question (RQ 2) to understand how to select the best ML approach in supporting users to find evidence. We found that a simple ED model that

adapts to an individual topic often outperforms a well generalising model that was trained out-of-domain. This chapter is based on Stahlhut (2019b) and contains contributions ③, ④, and ⑤.

Chapter 6 Machine Learning for Evidence Linking:

In this chapter we investigate which ML approach is most appropriate when aiming at supporting users in linking pieces of evidence to self-defined hypotheses (RQ 3). We found that a well generalising but out-of-domain trained EL model outperforms models which have been trained or fine-tuned on the user-specific data. We also found that using weakly labelled links to augment the user-created links results in a dataset which can confuse a machine learner. This chapter contains work from Stahlhut (2018) and parts of contribution ⑦.

Chapter 7 Interactive Evidence Detection:

We combine the results of the previous chapters to evaluate whether or not learning to find evidence is beneficial to actual users (RQ 4). We do so by extending the previously described annotation tool EDoHa with interactively trained ED and EL models. We then evaluate the effects of the suggestions done by EDoHa qualitatively with students and expert users. This chapter is based on Stahlhut (2019a) and contains parts of contribution ⑥ and ⑦.

Chapter 8 Conclusion:

This chapter summarises the contributions of this thesis and lies out future directions of research.

Chapter 2

Research Methodologies and Evidence

Before we can address our research questions, we must first understand what evidence and hypotheses are, how they are used, and lay out the conceptual framework of this thesis. We use the term conceptual framework rather than theoretical framework to avoid confusion between different meanings of the term *theory*. Given that this thesis aims at a degree in the natural sciences, any mention of the term theory would imply the meaning of a scientific theory, which is not in line with the purpose of this chapter; exceptions here are parts of names of specific methodologies, such as grounded theory. Similarly varied is the idea of evidence in different disciplines. Whereas a scientist sees evidence mostly as something produced when running experiments, researchers in history or sociology produce their evidence in a different manner. These differences also depend on the general epistemological approach, which falls into one of two categories: the *Hypothetico-Deductive (HD)* and *inductive* approaches.

In the HD approach, which research in NLP follows, the hypotheses are defined before the collection of evidence begins. This means formulating a hypothesis regarding how one can address a particular problem, such as ED, and then conducting experiments that create evidence that can validate or invalidate this hypothesis. These experiments need to be carefully designed to take into account any relevant influence that might interfere with the results, such as the influence of stochastic components, or noise in the data. The result of these experiments will then be evidence that either supports or contradicts the hypotheses.

Following an inductive approach, on the other hand, means collecting the evidence first, and then building hypotheses based on it (Andersen and Hepburn, 2016). For example, research in the field of history follows more of an inductive than an HD approach. Instead of formulating a hypothesis in the beginning and evaluating it on a held-out dataset, the researcher uses early working hypotheses to identify potential sources. However, the working hypotheses are not fixed; they are developed over time, which is more akin to an inductive rather than HD approach. Additionally, the goal of historical research is not only to document *what* happened but *how* it happened and to sometimes try to explain *why* it happened. This influences what kind of hypotheses a researcher formulates, and in turn, what is evidential and what is not. Some research methods in sociology, such as *grounded theory* (Strauss and Corbin, 1994), also follow an inductive research approach.

However, when looking at research practice, the HD approach is not followed to the

letter (Hanson, 1983); it often uses inductive approaches to bootstrap hypotheses. To better understand the challenges arising in supporting researchers, we take a look at the general approaches of Hypothesis Validation (HV), here with a focus on the sciences, particularly in NLP and in historical and sociological research.

2.1 Evidence in Philosophy of Science

When searching for definitions of scientific methods, one of the most well-known and understandable ones is from Feynman (1985, p. 156):

In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is - if it disagrees with experiment it is wrong. That is all there is to it.

Although this definition is a simplified description of the falsification approach (Andersen and Hepburn, 2016), it clearly states that a single contradicting piece of evidence must lead to the rejection of the hypothesis; agreeing evidence only shows that the hypothesis is not wrong (Feynman, 1985, p. 157). However, this direct rejection does not necessarily follow in practice. For instance, the orbital motion of the planet Uranus does not agree with Newtonian physics, leading astronomers to the *ad hoc* hypothesis (Andersen and Hepburn, 2016) that there is another planet influencing its orbit, this eventually led to the discovery of Neptune. The reason for the creation of this *ad hoc* hypothesis rather than the immediate rejection of Newton’s law of gravity is the “incredible success” (Popper, 2005, p. 38) of Newton’s law of gravity, or in other words, the enormous amount of supporting evidence. Therefore, we have to consider how well a hypothesis is supported.

Evaluating how well a hypothesis is supported is greatly debated within the philosophy of science. Some approaches are particularly relevant to research in Claim Verification (CV), because they contain fundamental arguments and problems.

A naïve way of validating or confirming a hypothesis is to reason that if all available evidence is supporting the hypothesis, then it is also confirmed because it currently can be. However, as Hempel (1983) pointed out, a particular hypothesis can be represented in different ways, and evidence must support all representations. They pointed out that the hypothesis “*All ravens are black*” is identical to the hypothesis “*No raven is not black*”¹. Although both hypotheses can be invalidated by a single raven that is not black, it does not mean that the same piece of evidence will support both hypotheses. This can be illustrated better when considering the hypotheses as statements in predicate logic.

$$\text{All Ravens are Black} = \forall x : R(x) \implies B(x) \quad (2.1)$$

$$\text{No Raven is not Black} = \forall x : \neg B(x) \implies \neg R(x). \quad (2.2)$$

Considering $a \implies b \equiv \neg a \wedge b$, that is if a is true, then b must be true for the entire expression to be true, allows us to define the options for relevant evidence; if a is not

¹This hypothesis is evidently false, given the existence of white ravens, which are not black. However, it is still useful for illustrative purposes.

true, then the expression is already true and the value of b is irrelevant. Therefore, when searching for evidence in the case of (2.1), anything that is not a raven ($\neg R(x)$) already evaluates to true and is uninteresting. This means, creating evidence would start by collecting as many ravens as possible and to ascertain whether any of them are not black. For (2.2), on the other hand, if x is black, then the expression again evaluates to true, and therefore the evidence would be created by collecting as many non-black birds as possible and by making sure that neither of them is a raven. For instance, a white swan, although not evidence of the first representation, is still evidence supporting the second, because it is not black but also not a raven.

Carnap (1983) extended the notion of confirming evidence by distinguishing between confirmation as the conclusion that a hypothesis is valid and confirmation in a piece of evidence supporting the hypothesis. The latter definition then allows for defining *degrees of confirmation* by declaring it as a probability distribution function for which a new piece of evidence increases the degree of confirmation. In turn, this allows for comparing two hypotheses and reasoning about which one is better supported than the other. Salmon (1983) criticised treating the degree of confirmation as a probability, because contradicting evidence would have a negative probability, which is a value a probability cannot have. Still, they agreed that treating the degree of confirmation similarly to a probability would enable Bayesian reasoning. Salmon (1983) also added that a piece of evidence that supports two hypotheses might not necessarily support their conjunction. This can be the case if the two hypotheses are mutually exclusive.

Still, it might be possible that evidence supports multiple hypotheses, and this opens to the question of which of these hypotheses to select. A common heuristic is the application of Ockham's razor, which states that among two hypotheses, the preference should be on the hypothesis that requires a smaller ontology (Baker, 2016). However, such a heuristic does not allow for a formal definition. Goodman (1983) suggested distinguishing between hypotheses based on the idea of *projection*. The projection of a hypothesis is to inductively reason from available evidence that future evidence also supports a hypothesis. They defined the projectibility of a hypothesis as follows (Cohnitz and Rossberg, 2019):

1. A hypothesis is *projectible* if and only if it is supported, not exhausted, not violated, and all conflicting hypotheses are overridden.
2. A hypothesis is *unprojectible* if it is either unsupported, exhausted, violated, or overridden.
3. A hypothesis is *nonprojectible* if and only if a conflicting hypothesis and itself are supported, not violated, not exhausted, and not overridden.

Supported in this context means that there is evidence supporting the hypothesis and violated means that there is contradicting evidence. An exhausted hypothesis is one for which there cannot be any more evidence that supports or contradicts it. A hypothesis h overrides another hypothesis k if all predicates of $h = \{p, q\}$ are supported, whereas not all predicates of $k = \{p, r, z\}$ are supported by evidence. It is possible that the same evidence can support predicates from either hypothesis. Goodman (1983) illustrated this with two competing hypotheses. First, “All emeralds are green” and second, “All emeralds are *grue*”, where *grue* is the predicate that it is green until time t and blue afterwards. Because all pieces of evidence, collected before time t , that support the first hypothesis also support the second one, both hypotheses are supported. However, until we reach t , the predicate that an item observed after the time t is blue, is not supported. Therefore, the first hypothesis “All emeralds are green” overrides the hypothesis “All emeralds are *grue*” until time t is reached. Afterwards, each newly collected piece of evidence will either contradict the first

or the second hypothesis, invalidating one of them. Figure 2.1 illustrates the difference between the first and second hypotheses.

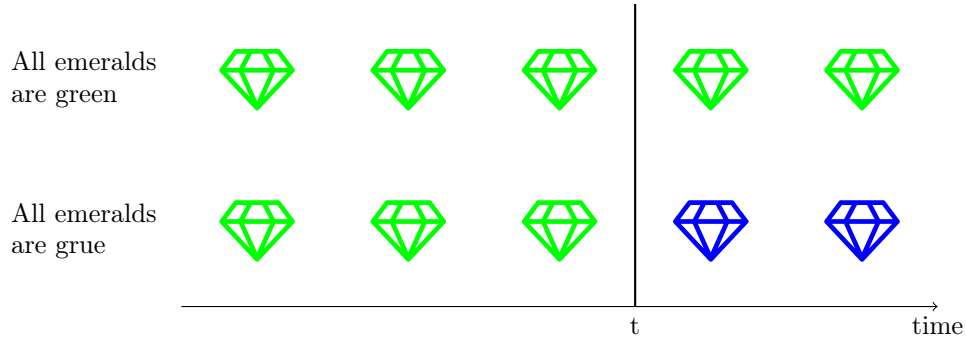


Figure 2.1: Illustration of the differences between the hypotheses that all emeralds are green and that all emeralds are grue.

This also shows the importance of *novel evidence* because a hypothesis must also be confirmed by future evidence. This means that both hypotheses regarding the colour of emeralds must not only be supported by the evidence used to formulate them but also by the evidence created afterwards and at all points in time. Any emerald examined before t will be green, thereby supporting both hypotheses. Only after t it is possible to fully evaluate the second hypothesis and determine which one is more likely to be true.

2.2 Evidence in Natural Language Processing

In NLP and ML in general, researchers take great care to set up experiments that result in the best available evidence. The result of these experiments are predictions on previously unseen data. These predictions are then evaluated using different metrics. In the case of supervised learning, which means that there is labelled data, i.e. for each testing datum it is known what the target result is, metrics can be used to compare the predicted result with the target result.

2.2.1 Evaluation Setups

If a researcher in NLP wishes to investigate whether a new method or model works better than an old one, they approach this through experiments. The purpose of these experiments is twofold. First, to simulate how a model would behave in real world scenarios without having to bring the world into the laboratory. Second, to have consistent results for replication while also having results that are comparable with previous work.

Generally, there are two different kinds of experiments: *intrinsic* and *extrinsic* (Resnik and J. Lin, 2010). Intrinsic experiments rely on data alone, which form the *ground truth* against which a model's predictions are compared. The combination of the model's prediction and ground truth then forms the evidence that we interpret using different metrics. If the hypothesis contains any reference point, such as a baseline or some other competing model, the predictions of the baseline or competing model become part of the evidence as well. Although these experiments can give quantitative results that allow for distinguishing good from bad models, intrinsic experiments suffer from two problems. First, if the evaluation metric has flaws, it might result in sub-optimal models that, despite reaching high performance scores, do not actually work better than those with lower scores (Dorr et al., 2005). Second, even if the metric does not have such a flaw, it is only a substitute

to decide which model to use in real-world situations. Therefore, to assess the quality of a model, researchers in NLP use extrinsic experiments in which a particular model is used to perform its task in a wider context. The evaluation can even be done via real human users, who then give feedback on the model's predictive quality. Evidence in an extrinsic evaluation is more varied than in an intrinsic experiment. It can be similar to the intrinsic evaluation in that it consists of combinations of ground truth and predictions, but also of feedback provided by users.

Intrinsic Evaluations

The setup of an intrinsic experiment depends not only on the model to be evaluated, but also on the amount of available data. If the amount of data is plentiful, we can split it into three parts: *training*, *development* or *validation*, and *testing*. The training data are used to train or fit the model, and the testing data are used to produce the evidence showing how well it performs on the task. It is of vital importance that at the time of the evaluation, the testing data are unknown to the model. Otherwise, for instance, if parts of the testing data are also in the training data, the model will have seen it and its predictions would not be realistic. Being able to make perfect predictions on this testing data overlapping with the training data would lead to an overestimation of the quality of the model because the model's predictions would be too good.

Development data are a third held-out dataset used to simulate the behaviour on the testing data. It generally serves two purposes: to do *hyper-parameter optimisation* and allow for *early stopping*. Hyper-parameter optimisation is a search through all the hyper-parameters, such as the features used as an input to the model, the learning rate, or number of updates for the model. Early stopping is based on the idea that when trained for too long, a model will start to *overfit* on the training data, which means that the model's predictions are becoming better on the training data but worse on other data. The development dataset is then used periodically to compare the current quality of the model with the previous one. If the model has not improved when evaluated on the development data after a pre-defined number of evaluations, the training will stop. Because the testing data are distinct from the development data, the model will also not overfit on the testing data. If the testing data were used to determine when to stop, it would lead to a model that is overfitted on the testing data, which means its evaluation would be invalid. Figure 2.2 illustrates the split into training, development, and testing data and the purpose of each split.

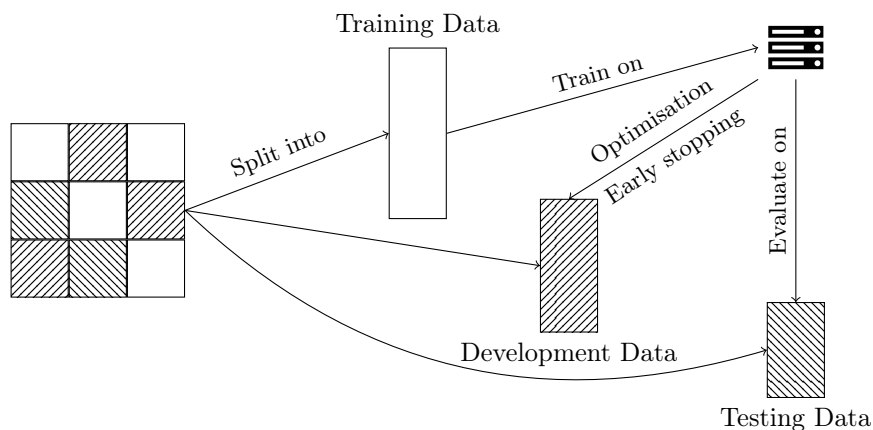


Figure 2.2: Illustration of a dataset split into training, develop, and testing data.

Splitting a dataset into three parts requires that the amount of data is plentiful. If this is not the case, splitting the data into three or even two parts can mean that the testing data is not realistic because it doesn't cover all possibilities that would arise in real-world situations. In such situations, a viable approach of splitting the data is to perform a *cross-validation*. In a cross-validation, each datum is used in both training and testing, but not at the same time. The dataset is split into folds of training and testing splits. For instance in a five fold cross-validation, the data is randomly split into 80% for training and 20% for testing. This is then repeated four times, building the five folds. In each fold, a different 20% of the data is used for testing, meaning that each datum is used exactly once for testing. The predictions on the testing data from each fold can then be aggregated and compared against the target data for the entire dataset. Another cross-validation approach does not use random splits, but leaves out a particular set of data; for instance *leave-one-file-out* uses one file as test data and the other as training data and after the first fold uses another file as test data. Figure 2.3 shows a three-fold cross-validation with the split into training and testing data.

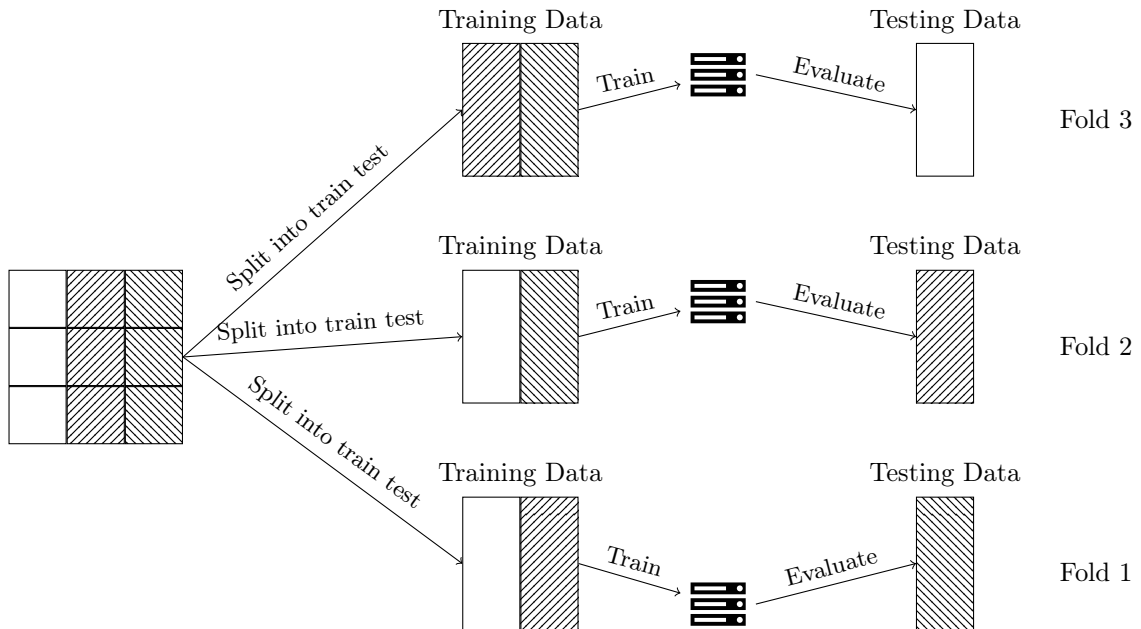


Figure 2.3: Illustration of a three-fold cross validation.

Oftentimes, there are additional influences that result in unreliable results. For instance, artificial neural networks use random values to initialise the weight matrices. This then results in variations of the initial behaviour of the models leading to different final states in the weight matrices and variations of the quality of the models. A common approach in addressing this issue is to repeat the experiment multiple times and use statistical methods in the evaluation (Reimers and Gurevych, 2018). Still, this approach does not remove the uncertainty induced by the non-deterministic nature of the experiment. Therefore, a common approach is to make the experiment deterministic and use a pseudo-random generator with a *randomisation seed*, which is generally seen as being good enough. In this way, we can still evaluate the different initial configurations, hence allowing for evaluation of the robustness with respect to these modifications, while still keeping the experiment repeatable.²

²By repeatability, we mean getting the same results under identical conditions, i.e. with the same software and on the same system. Repeatability is closely related to *reproducibility* and *replicability*, which

In NLP, one goal is to solve a particular task, e.g. evidence detection, argument mining, or semantic role labelling, with computational methods. However, no feasible experiment can be exhaustive; there might be other topics that have not been covered in the dataset or particular sequences of words. This means that an experiment on a single dataset is insufficient when it comes to claiming to have solved and possibly even advanced research regarding a task. Therefore, to advance research in a particular task using more than one dataset results in more informative evidence because it shows that a method is independent from a particular dataset. Furthermore, using datasets for different tasks results in evidence about the robustness of the evaluated method.

Extrinsic Evaluations

Although intrinsic evaluations are more common in NLP, oftentimes, it is quite relevant how a particular method or model performs in the context of actual or potential downstream tasks and users. For instance, Mollá and Hutchinson (2003) first used an intrinsic evaluation of two parsers and then a downstream task, namely *answer extraction*, to evaluate the usefulness of either parser. Another common extrinsic evaluation considers the use of actual human users, for example, in the evaluation of the argument search engine ArgumenText (Stab et al., 2018a).

Depending on the purpose of an extrinsic evaluation with actual users, the experiment can be conducted with few or as many users as possible. If the purpose is to validate the results of an intrinsic experiment, then the general requirements in the number of users and control would be lower than if the extrinsic experiment yields the primary evidence. For instance, Gao, Meyer, and Gurevych (2018) used seven users to evaluate the quality of their interactively trained summarisation system that they had also evaluated intrinsically. Głowacka et al. (2013), on the other hand, used 20 participants in a double-blind, controlled evaluation of their exploratory ranking learning system. Here, double-blind means that neither the participants, nor the experts who introduced the participants to the task at hand and evaluated the results knew, which ranking model was used, nor which student found which document. However, running such a controlled and double-blind experiment requires many users which increases the cost and effort. This means that it is not always possible to run a controlled experiment with a large number of users, instead researchers use qualitative methods, such as questionnaires, interviews, and general observations (e.g. Stab et al., 2018a).

2.2.2 Metrics

The primary influence on the choice of metric is what kind of task we evaluate a model. If the goal for a particular task is to predict a specific mathematical function, e.g. via linear regression, we use metrics that compare the continuous difference between the predicted and target result, such as *mean squared error* or *absolute mean error*. In ranking tasks where a list of items is supposed to be sorted by a learned criterion, e.g. results in an information retrieval system, the most common metrics would be the *mean reciprocal rank*, which returns the average of the inverse of the rank of the target item. For classification, there are four common metrics, namely *accuracy*, *precision*, *recall*, and *F1 score*, which are used to evaluate the quality or performance of a classification model.

Accuracy is the simplest and easiest to understand metric. It is defined as the ratio of correct predictions, or true positives tp and true negatives tn , over all positive P and

research in NLP strives to achieve but that have varying meanings (Plessner, 2018).

negative N samples. It is formally defined as

$$\text{Acc} = \frac{tp + tn}{P + N}. \quad (2.3)$$

Although easy to understand, it can be hard to interpret, particularly when a dataset is not balanced. If the majority class is 10 times more abundant than the minority class and the predictions are all on the majority class, then the accuracy is still 90%.

Precision and recall are two metrics that, combined, attempt to remedy the shortcomings of the accuracy metric. The precision of a particular class measures how many of the predictions were done correctly, and recall measures how many of the existing predictions were done. Precision is defined as the ratio of true positives over all predictions, or true positives and false positives fp and recall as the ratio of true positives over true positives and false negatives.

$$P = \frac{tp}{tp + fp} \quad (2.4)$$

$$R = \frac{tp}{tp + fn}. \quad (2.5)$$

Either metric alone is flawed; if a model classifies only a single item as belonging to a class and this classification is correct, then it has a precision of 100%, and if it classifies everything as belonging to the class, then the recall is 100%. Only together do these metrics provide full information about the quality of a model regarding a particular class. A model might have a very high precision, but if its recall is very low, this means that it only classifies a few instances. Depending on the purpose, it might be desirable to have a high precision or a high recall model. A high precision model can be used on large amounts of data without creating many false positives, and a high recall model might be desirable when it is important to get as many instances of the prediction class as possible.

When comparing multiple different models with one another, using two metrics that describe the quality of either is inconvenient, which is why the most common metric used in classification tasks in NLP is the F1 score. The F1 score is the *harmonic mean* of both the precision and recall,

$$F1 = 2 \times \frac{P \times R}{P + R}. \quad (2.6)$$

Using this metric allows researchers to compare different models with one another without being deceived by either the shortcomings of precision or recall.

So far, we have considered precision, recall, and F1 score only regarding the quality of a single class. However, many tasks in NLP use multiple classes, e.g. supporting and contradicting argument and neutral statements. In this case, one can list the F1 score of each class or build an average across all classes. When averaging, there are two methods: One is the *macro-average*, which is the arithmetic mean across the scores of all classes, meaning if the F1 score on the majority class is 80% and the F1 score on the minority class is 40%, the macro-averaged score is 60%. The second kind of average is the *micro-average*. This average does take the class imbalance into account. For instance, in the previous example, if the majority class consists of 90% of all samples, the micro-averaged score will be 76%. Commonly, the macro-average is reported because it is less affected by class imbalances than the micro-average (Forman and Scholz, 2010).

2.3 Evidence in History

To understand the meaning of evidence in historical research, we must first understand what historical research is. Little (2017) defined historians as follows:

In short, historians conceptualize, describe, contextualize, explain, and interpret events and circumstances of the past. They sketch out ways of representing the complex activities and events of the past; they explain and interpret significant outcomes; and they base their findings on evidence in the present that bears upon facts about the past. [...] [H]istorical statements depend ultimately upon factual inquiry and theoretical reasoning. Ultimately, the historian's task is to shed light on the what, why, and how of the past, based on inferences from the evidence of the present.

This means the goal of historical research is not only to create a chronicle of past events, answering the question *what happened*, but also to explain the circumstances and answer questions, such as *how did it happen* and *why did it happen*. This also means that historians need multiple different kinds of evidence, one for each question. Evidence regarding *what* questions can sometimes be extracted directly or indirectly from the sources, e.g. a report states that a street had been built or a source states that someone travelled on a particular street, meaning that the street had to be built earlier. Of course, this assumes that the source is trustworthy, which also must be established based on the sources' context. Some *What* questions are harder to answer, e.g. "*What did Pius XII know about the holocaust?*". Questions of the type *how* and *why* require knowledge about the circumstances and human and social behaviour.

Although the interpretation and contextualisation of sources is a necessary part of historical research in building a narrative, they cannot contradict credible sources. This also means that narratives can change as new sources are discovered and new evidence is unveiled. This principle is known as *Vetorecht der Quellen*, first introduced by Reinhard Koselleck (Jordan, 2010), and sets boundaries to interpretations. An interpretation that directly conflicts established and trusted sources must be wrong. Here, historical research shares the idea of falsifiability with other fields of research and sciences.

What sources are and where to get the evidence from can vary greatly in historical research. For instance, Daston (1991) relied on historical publications and pictorial evidence from paintings to investigate the history of neutral facts and evidence. Schaffer (1992), on the other hand, relied on publications, as well as autobiographies and private letters to investigate how the requirement for trustworthy evidence changed in the late 18th and early 19th centuries. This shows that although text is the primary source of evidence in historical research, it is not the only one.

Furthermore, in historical research, there are different *schools of thought* (Lloyd, 2008) used in the interpretation and contextualisation of the evidence. Two examples of different schools of thought are *historical materialism* and *subaltern studies*.

Historical materialism This school of thought, which was initially developed by Karl Marx and often referred to as Marxist historiography, sets increased productivity and the resulting conflict between social classes as the "driving force of history" (Wolff, 2017). If an economic system can no longer increase its productivity, it needs to be replaced by one that does. For instance, a ruling class clinging onto their power by repressing other classes hampers the increase in productivity. Hence, the lower classes start to rebel until they overthrow the ruling classes and can build a more productive society. This repeats itself until the workers are no longer suppressed, meaning that they are in charge of the society in which they live.

Subaltern studies This school of thought criticises other historiographic schools, e.g. nationalistic history, by pointing out their focus on the elites of a region or country. Instead

of focussing on the accomplishments of the elites, subaltern studies place the subordinate classes, or *subalterns*, into historical focus (Chakrabarty, 2000). This view draws from Marxist historiography in that it views power relations and class-struggle as important but sees them as independent from “universalist histories of capital” (Chakrabarty, 2000). Instead, it focusses on the connection between power and historical knowledge.

The difference between these schools of thought can be illustrated by analysing the history of India and its development from a British colony into an independent nation. Marxist historians would focus on the interaction between the different social classes in the context of capital. Subaltern studies, on the other hand, criticise the historiographical materialism as being Eurocentric, meaning that it assumes every other region to work similarly and ignoring many differences. Therefore, it would focus primarily on the subalterns, giving them a voice in their history.

These different intellectual frameworks also change how a historian would see their sources. Not only would two historians from different historiographic schools formulate different hypotheses, they would also see the sources with different eyes, hence seeing completely different evidence.

2.4 Evidence in Sociology

The American Sociological Association (2008) defined what sociologists study as follows:

Sociologists study all things human, from the interactions between two people to the complex relationships between nations or multinational corporations. While sociology assumes that human actions are patterned, individuals still have room for choices. Becoming aware of the social processes that influence the way humans think, feel, and behave plus having the will to act can help individuals to shape the social forces they face.

With its focus on humans, sociology can take many shapes and investigate many different parts of human society. This includes the complexity of human societies, how societies acquire knowledge, or sociology of knowledge, and general criticism of human interactions. For instance, if a sociologist would study anti-genetically modified organisms activism, they would focus on the actions and interactions of the activists, not on the validity of the arguments. If an activist proclaims, “*They do not want genes in their food*”, a sociologist does not investigate the possibility of creating food without genes or whether this statement is meant as a shorthand for “*They do not want modified genes in their food*”. A sociologist’s interest lies in how this view is distributed in the activists’ society and how it affects their actions.

Similar to historical research, different schools of thought can be found in sociology. Some of these schools are shared between historians and sociologists. For instance, both Marxism and subaltern studies can be found in both historical research and sociology. However, many differences in sociology appear at a methodological level. For instance, there is a great divide between using *quantitative* or *qualitative* methods (Baur and Knoblauch, 2018). Even within qualitative methods, the general approach differs. Two examples are *Critical theory* and *Grounded theory*.

Critical theory of society Geuss (1981, p. 55f) stated that in contrast to scientific theories, which aim at describing the environment, critical theories have the goal of “emancipation and enlightenment”. This means that their goal is not to describe and explain a phenomenon, but to enable an agent to identify the “true interests” (Geuss, 1981, p. 2) of

the agent and to free them from otherwise invisible coercion. A central method of critical theory of society is the criticism of ideology, which limits the perception and freedom of the members of that society. A second difference between scientific theories and critical theories lies in their epistemology. Whereas scientific theories are evaluated empirically with evidence created in observations and experiments, critical theories are evaluated in a reflective manner. This means that the evidence in a critical theory is the experience of the reflection itself.

Grounded theory Similar to critical theory, grounded theory is not a theory that explains a phenomenon, but is instead a methodology that allows for building a theory *grounded* in data (Strauss and Corbin, 1994). It is an inductive research method in which the goal is to create a theory based on the available data while also conducting the research and developing a theory. Theories in this context “are interpretations made from given perspectives” (Strauss and Corbin, 1994, p. 279) based on the plausible relationships between individuals or sets of concepts. The concepts are developed in analysing the source material, which can have many different forms, such as transcriptions of interviews, historical documents, or videos. Afterwards, the concepts are grouped together in a *conditional matrix*³ that visualises the similarity between the concepts. The researcher can then go back to the source material and continue to develop and improve the concepts. Therefore, the theory is not rigid, but developing continuously over time. However, because grounded theory is a method rather than an intellectual framework, its users can still be working within a particular school of thought.

From these two examples, we can conclude that there is not a single unified approach within the field of sociology and their use of evidence. Even what is considered evidence varies greatly. A critical theorist perceives the experience of reflecting and criticising as evidence, whereas a researcher using grounded theory would rely on the statements in the source material. The flexibility of grounded theory, enabling it to be used by researchers from many different schools of thought, also results in two researchers using the same method, but disagreeing completely when it comes to evidence. Furthermore, the fluidity of the theorist’s understanding and goal also poses additional challenges.

2.5 Chapter Conclusion

In this chapter, we developed the conceptual foundation of this thesis. We investigated the role evidence plays when (in)validating hypotheses in different academic disciplines. We described that a piece of evidence is a fact used in an argument and in giving credibility to hypotheses. We also looked at the different approaches from the philosophy of science, hence allow for comparing hypotheses regarding their degree of confirmation, e.g. the concept of *projectibility*, and found that novel evidence can change drastically which hypothesis is better supported. For instance, the hypotheses that “*all emeralds are green*” and “*all emeralds are grue*” are equally supported until time t , even though the first one is more projectable. A newly examined blue emerald changes this drastically, because it contradicts the first hypothesis and supports the second one, meaning that the second one is then the better-confirmed hypothesis.

We also described the epistemology of NLP and ML and how different experimental setups can be used to create evidence to evaluate a variety of hypotheses. Although an intrinsic evaluation based on gold-standard data is generally preferred, it is neither the only experimental setup, nor always the most appropriate one. Quite often, an extrinsic

³A common approach is to use circles around concepts rather than a mathematical matrix.

evaluation, in which real users' feedback is used, gives more informative evidence. We also described how the evidence created in intrinsic experiments is then interpreted using different metrics. In particular, we looked at how to aggregate metrics for different purposes, such as precision and recall to an F1 score, and how to aggregate repeated experiments required by the non-deterministic nature of some ML methods.

In the final parts of this chapter, we looked at the relevant research disciplines; what their research focus is and how they use evidence. We focussed on the research disciplines from which we extracted our use cases, namely *history* and *sociology*. In this analysis, we found that what is evidence depends heavily not only on the academic discipline, but to a large extent on the particular school of thought an individual researcher falls into. For instance, in history, although the schools of historical materialism and subaltern studies both focus on power relations between different social classes, they do not overlap completely. Whereas historical materialism always considers the productivity of a society, subaltern studies criticise this as being Eurocentric and instead focusses on the previously unheard voices or subalterns and the factors influencing them. These factors can be economic, but are often cultural, e.g. in the form of racism, or even legal in nature. Sociology shows an even larger difference, in particular with schools of thought such as critical theory, in which evidence is the experience of reflection. On the other hand, sociology uses methods such as grounded theory in which the researcher develops a theory based on data, which means the evidence supporting the theory is independent from particular schools of thought as long as the method is applicable. Furthermore, in grounded theoretical approaches, the theory is continuously changing and, therefore, very fluid.

Researchers in history and sociology cannot be supported in finding evidence and validating hypotheses by creating a single gold-standard dataset that abstracts the entire task. Any approach in creating a dataset must at least take into account the existence of different schools of thought, that evidence varies between them, and especially, what a piece of evidence means. The fluidity of some approaches, such as grounded theory, pose even more challenges, because the users are not focussing on a fixed long-term goal, but change their goals continuously as their understanding grows.

Chapter 3

Automatic Support for Evidence Detection and Hypothesis Validation

Supporting researchers in history and sociology in finding evidence and validating their hypotheses touches on several concepts, tasks, and approaches within NLP and ML. The related task of Claim Verification (CV) is gaining popularity because of the increased focus on fighting fake news. This increased focus has led to considerable improvements in using automated methods to detect not only fake news, but also to validate claims in general. Although a hypothesis is always a claim, it is unclear how well general purpose CV methods and datasets can be applied when dealing with research hypotheses. For instance, when creating a CV dataset, the claims must be verifiable from the available sources. As we discussed in chapter 2, what kind of hypothesis a researcher develops can depend heavily on the researcher’s intellectual framework. Furthermore, methodologies such as grounded theory are fluid in nature and many ED and CV approaches rely on static models that do not consider changes in the user’s goal.

Besides ED and CV, our research touches on other tasks and approaches. The task of ED is often viewed as a sub-task of Argument Mining (AM), in which the hypothesis-evidence pair is treated as a *claim* and *premise*. A piece of evidence is a premise supporting or contradicting the hypothesis. The difference between a piece of evidence and a general purpose premise lies in the factuality of the evidence, whereas general purpose premise can be true but not factual statement. For instance, in the argument “*The idea of the end justifying the means is bad for society because it can lead to unnecessary deaths in times of crisis*” the premise “*it can lead to unnecessary deaths in times of crisis*” is generally seen as true, but it is not a factual statement, so that is not a piece of evidence. In analysing the existing work in ED, CV, and AM, we consider their applicability in supporting researchers in history and sociology. In particular, the applicability to research, as in conceptualisation and contextualisation, focusses on different schools of thought and fluidity.

Another related task is *interactive machine learning*, because we aim at learning interactively from a user. In interactive machine learning, there is generally a *learner* trying to learn a task in interaction with a human or another kind of *oracle*.

Because the current thesis aims at supporting researchers in history and sociology, it is also closely related to the field of *digital humanities*. Digital humanities is a research discipline that combines computational and humanities approaches with one another to

investigate different kinds of phenomenon. Although this thesis makes contributions to research methodologies in history and sociology, its primary focus lies in ED and NLP, making it related to digital humanities, but not part of it.

3.1 Current Research in Evidence Detection

Research into ED can be separated based on *datasets* and *methods*; either one reflects the large variety of the evidence itself. The aim of the ED dataset can vary from debating (Rinott et al., 2015) to medical research (Mayer et al., 2018). Whereas some researchers built datasets for ED from Wikipedia articles (e.g. Aharoni et al., 2014), others used news editorials (Al-Khatib et al., 2016). Similar variation can be found in the methods. Although at first, ED seems like a classification problem, it can also be treated as a ranking problem, in which the goal is to rank a group of candidate pieces of evidence with respect to a given claim. Furthermore, some research focusses on parts of the ED problem, such as classifying different types of evidence, e.g. *expert opinions* and *study data*.

3.1.1 Datasets

The published datasets vary greatly in their aim, data source, and size. Table 3.1 summarises the datasets analysed in the present thesis.

Author (Label)	Aim	Types	Size
Aharoni et al. (2014) (ED-ACL-2014)	Debating	Claims, Anecdote, Expert Opinion, and Study Data	12 topics with evidence, 386 documents, 1387 claims, 1291 pieces of evidence
Rinott et al. (2015) (ED-EMNLP-2015)	Debating	Claims, Anecdote, Expert Opinion, and Study Data	58 topics, 1298 documents, 2295 claims, 4692 pieces of evidence
Al-Khatib et al. (2016) (Editorial-2016)	Persuasion	Assumption, Anecdote, Statistics, Testimony, Common ground, other	300 documents, 9792 assumptions, 2603 anecdotes, 1089 testimonies, 421 statistics, 241 common ground, 167 other
Mayer et al. (2018) (PubMed-2018)	Medical Decision Making	Claim and Evidence	4 Topics, 169 abstracts, 279 claims, 697 pieces of evidence
Shnarch et al. (2018) (ED-ACL-2018)	Debating	Topic and Evidence	118 topics, 5783 sentences, 2182 pieces of evidence

Table 3.1: Overview of the existing datasets for ED.

ED-ACL-2014 Aharoni et al. (2014) first extracted 33 topics from the debate portal idebate¹ and then used five expert annotators to extract claims and later evidence

¹<https://idebate.org/>

from Wikipedia articles. In the first step, the annotators searched for related articles on Wikipedia, and afterwards, five annotators searched for claims within the articles that were then cross-examined by the other annotators. They then used majority voting for each claim before including the claim into the dataset. The next step in the dataset creation was to select 12 topics for which to annotate pieces of evidence related to the previously annotated claims. Much like to the claim annotation, they used five annotators to search for evidence related to each confirmed claim within the same document, in which the claim was found. As before, they used a cross-examination approach with majority voting to include evidence into the dataset. They also included evidence type annotation into the dataset, which means that a piece of evidence can be a study data, expert opinion, or an anecdote. The evidence types were not aggregated via majority voting, but a different, more complicated way of achieving consensus. The dataset consists of 1.3k claims, 386 Wikipedia articles and $\approx 1.3k$ pieces of evidence.

ED-EMNLP-2015 The dataset created by Rinott et al. (2015) was constructed in a similar fashion as the ED-ACL-2014. However, they extracted 58 topics from idebate and 1298 Wikipedia articles, 547 of which contain pieces of evidence. They split it on a topic level into train/dev and train/test; the train/dev split can be used for hyper-parameter tuning.

Editorial-2016 The dataset presented by Al-Khatib et al. (2016) consists of 100 editorials extracted from Al Jazeera, Fox News, and The Guardian. The authors annotated six different types of evidence: *common ground*, *assumption*, *testimony*, *statistics*, *anecdote*, and *other*. The annotations consist mostly of assumptions and anecdotes, with the other classes appearing less than 10% each.

PubMed-2018 This dataset was released by Mayer et al. (2018) and is based on a corpus of 99 PICO (Population, Intervention, Control/Comparison, Outcome) annotated medical abstracts of randomised controlled trials (Trenta, Hunter, and Riedel, 2015) collected from the PubMed database² of the National Center for Biotechnology Information. Mayer et al. (2018) added discourse-level annotations to the abstracts in the form of *major claim*, *claim*, and *evidence*. The claims are the conclusions of the trial, with the major claim being the main conclusion. The pieces of evidence are the direct results leading to the conclusions.

ED-ACL-2018 This dataset was created by Shnarch et al. (2018) and consists of 118 topics extracted from debate portals and 5.7k sentences extracted from Wikipedia. The sentences were labelled as either evidence or not by ten crowdworkers each and aggregated via majority voting. The dataset was then split on the topic level with 83 topics and 4k topic sentence pairs used for training and 35 topics and 1.7k topic sentence pairs used for testing.

3.1.2 Methods

The methods developed for ED also vary greatly. The two most common approaches are to treat ED either as a *ranking* or as *classification* task. Figure 3.1 illustrates the difference between these approaches.

²<https://www.ncbi.nlm.nih.gov/pubmed/>

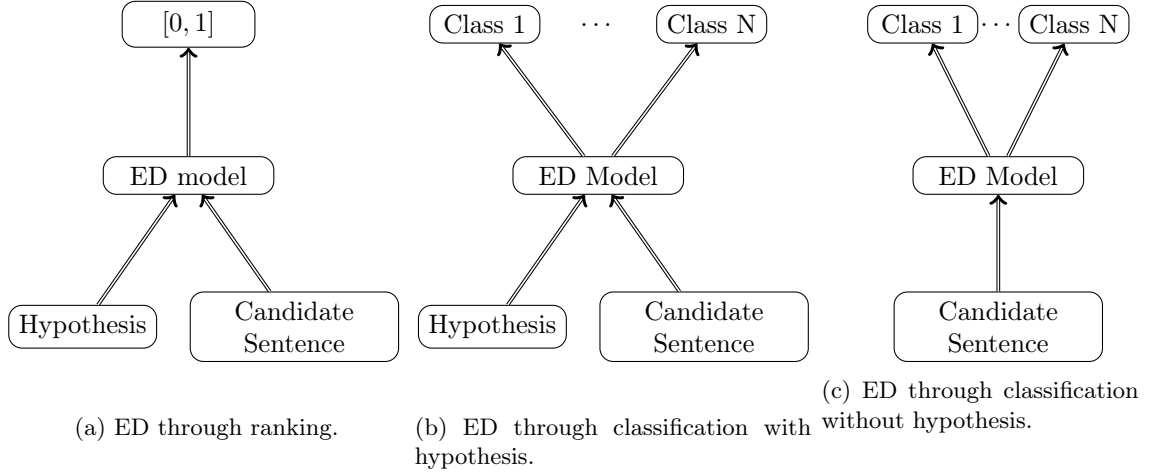


Figure 3.1: Different approaches to ED. a shows a ranking approach, whereas b shows a classification approach with hypothesis, and c shows a classification approach without access to the hypothesis.

Evidence Detection through Ranking

Rinott et al. (2015) used a pipeline of four components to rank evidence with respect to a user-given topic and a collection of claims. The first two components are context insensitive, meaning that for a given collection of articles, they return a collection of candidate pieces of evidence, irrespective of any potential relation to a given claim. The third component then ranks the candidates from the previous step regarding the given claims and the final component ranks the user provided claims, regarding their probability of pieces of evidence being present in the high ranking results of the previous step.

Hua and L. Wang (2017) treated finding supporting evidence for a given claim as a ranking problem. They used LambdaMART (Burgess, 2010) and a combination of similarity between the claim and candidate and composite features to rank sentences. Additionally, they used a feature engineering approach with a log-linear model to detect the type of evidence, i.e. study, factual, opinion, and reason. The features contain, among the basic features, such as n-grams and POS tags, values for sentiment, arousal, and discourse features.

Evidence Detection through Classification

MARGOT is a web service that can extract the claims and pieces of evidence based on both the ED-ACL-2014 and ED-EMNLP-2015 datasets published by Lippi and Torroni (2016). They treated both, claim detection and ED in a pipeline of two steps. The first step detects argumentative sentences, i.e. sentences that are either a claim or a piece of evidence, see figure 3.1c. The second step detects the specific boundaries of claims or pieces of evidence. Mayer et al. (2018) modified MARGOT to detect claims and evidence in the PubMed-2018 dataset. They later added evidence type detection, i.e. *comparative*, *side-effect*, *significance*, and *other*, and used a Support Vector Machine (SVM) with a constituency parse tree, a bag-of-words, and TF-IDF features and SubSet Tree Kernels (Collins and Duffy, 2002) to classify the evidence. Their approach distinguished the different classes well, except for the class *other*, and using their ED method before classifying their types reduced the macro-F1 score by about 5 pp.

A similar approach was used by Liga (2019), who also used tree kernels to classify pieces of evidence as either research results or testimony. They based their data on the

ED-ACL-2014 and Editorial-2016 datasets.

Research using the ED-ACL-2018 dataset follows the hypothesis-aware approach, as shown in figure 3.1b. Shnarch et al. (2018) combined the strongly labelled and weakly labelled data to investigate the benefit of first pre-training on the weakly labelled data and then blending weakly labelled and strongly labelled data. They found that pre-training and blending lead to better results with lower amounts of strongly labelled data. Reimers et al. (2019) used the same dataset and found that BERT outperformed the previously established state-of-the-art.

3.1.3 Relation to Research in History and Sociology

We aim to support researchers in history and sociology, which means we have to consider the applicability of the available datasets and methods in research in history or sociology.

Corpora

Neither dataset directly aims at supporting researchers in history or sociology, but some can be more supportive than others. Although all datasets contain both claims, as well as pieces of evidence, the PubMed-2018 and Editorial-2016 datasets were annotated with an approach different from others. The PubMed-2018 and Editorial-2016 datasets were annotated from the perspective of annotating the discourse. This means the creators of the datasets defined an annotation scheme and applied it to the texts at hand. If their annotation scheme contained claims, which it did in the case of the PubMed-2018 dataset, they labelled it accordingly. If the scheme contained a type also covering claims, e.g. *assumption*, they also labelled it as assumption. Although a claim and assumption are very similar and share a large overlap in instances, they are not the same. They also did not consider any content restriction. This means, in the case of the PubMed-2018 corpus, if the abstract contained the initial search keyword, i.e. the topic, but did not contain any argumentative component related to it, they annotated the entire abstract. In the case of the Editorial-2016 corpus, they randomly selected news editorials and annotated them. There was no preconceived topic to which the argumentative components were limited.

The other datasets, however, were created with an approach more similar to how a researcher would proceed. The annotators started from a controversial topic and then first selected suitable documents and before annotating topic-related claims and pieces of evidence. In the case of the ED-ACL-2014 and ED-EMNLP-2015 datasets, the annotators first labelled claims in documents. Afterwards, the annotators labelled pieces of evidence relevant to a claim in the same document in which the claim originated. The difference from the previously mentioned dataset is that the annotators used the topic as a filter for claims. If a document contained a claim that is not relevant to the topic, it would not be annotated. We conclude that the ED-ACL-2014 and ED-EMNLP-2015 datasets were created much like how a researcher approaches a new topic and so these datasets might be useful for user simulations to evaluate different methods.

Methods

Either approach, ranking as well as classification, has its benefits depending on the use case. Ranking seems to be the most applicable in cases in which a historian or sociologist would state their hypothesis beforehand and then search through a large collection of documents. If the ranking model would adapt to the individual user, this approach might be of great benefit. However, as discussed earlier, researchers in history and sociology do not necessarily formulate hypotheses beforehand, and especially for sociologists, their goal

is much more fluid. This could cause problems for a ranking model. Therefore, we treat both ED and EL as classification problems in this thesis.

Many of the presented methods depend heavily on complex pre-processing steps (e.g. Mayer et al., 2018; Liga, 2019). This also causes problems for researchers without knowledge about NLP or no background in software development. Furthermore, these pre-processing steps, e.g. parsing, vary greatly in quality between different languages (Schlichtkrull and Sogaard, 2017). Therefore, we focus on methods that require less pre-processing and fewer programming knowledge and are more stable across different languages.

3.2 Fact-Checking and Claim Verification

Claim Verification (CV), also known as *claim validation* or *fact verification*, is a task in which for a given claim, the validity or likelihood of validity is supposed to be determined automatically. It is closely related to *fake news detection* (Pomerleau and Rao, 2016) and deception detection (Rubin, Y. Chen, and Conroy, 2015). A common approach to address CV is to separate it into different tasks, such as *document retrieval*, *ED*, and *Textual Entailment (TE)* (Thorne et al., 2018b). Others integrate the last two tasks into one end-to-end system, e.g. DeClarE (Popat et al., 2018). A third common approach is to focus on individual tasks within CV, e.g. *document retrieval*, the previously covered ED, or the final decision-making task (Vlachos and Riedel, 2014). Table 3.2 summarises the datasets analysed in this thesis.

3.2.1 Datasets

PolitiFact-2014 Vlachos and Riedel (2014) created a dataset of 106 claims collected from *PolitiFact*³ together with the verdict as true, mostly true, half true, mostly false, and false.

Emergent Ferreira and Vlachos (2016) presented an approach for document-level stance detection with respect to a claim. They created a dataset of 300 claims and 2.5k articles with one of three labels, namely for, against, and observing.

LIAR W. Y. Wang (2017) also collected 12.8k from PolitiFact labelled by humans with one of six classes, namely pants-fire, false, barely-true, half-true, mostly-true, and true. In addition to the claim, they added the speaker of said claim and the context, e.g. whether it was made on social media or in a television broadcast. As additional background information, they included the speaker’s political affiliation, their home state, and credit history, i.e. a credibility score based on previous correct and incorrect statements.

Snopes-2017 Popat et al. (2017) extracted 4.8k claims from the fact checking website Snopes⁴ and Wikipedia⁵ and used an external search engine to find related articles by searching for the claim. They collected 130k web documents using this method and automatically determined their stance towards the claim. Additionally, they collected 100 known hoaxes and 57 fictitious people from Wikipedia with a known negative verdict and 2.8k and 1.5k web documents, respectively.

³<https://www.politifact.com/>

⁴<https://www.snopes.com/>

⁵https://en.wikipedia.org/wiki/Main_Page

Author (Label)	Aim	Types	Claim Labels	Size
Vlachos and Riedel (2014) (PolitiFact-2014)	Fact-Checking	Claim	True, Mostly True, Half True, False	106 claims
Ferreira and Vlachos (2016) (Emergent)	Stance Detection	Claim, Document Stance	For, Against, Observing	2595 documents, 300 claims
W. Y. Wang (2017) (LIAR)	Fake News Detection	Claim	True, Mostly True, Half True, Barely True, False, Pants Fire	12800 claims
Popat et al. (2017) (Snopes-2017)	Claim Credibility	Claim, Document	<i>Claim</i> : True, False; <i>Stance</i> : Sup- port, Refute	137628 documents, 5013 claims
Popat et al. (2018) (PolitiFact-2018)	Claim Credibility	Claim	True, False	87643 documents, 13525 claims
Thorne et al. (2018a) (FEVER)	Fact-Checking	Claim, Evidence	Supported, Refuted, Not Enough Info	185445 claims
Hanselowski et al. (2019) (Snopes-2019)	Fact-Checking	Claim, Evidence	True, Mostly True, Mostly False, False	14000 source documents, 6400 claims, 8000 pieces of evidence

Table 3.2: Overview of existing datasets for CV.

PolitiFact-2018 This is a dataset published by Popat et al. (2018) consisting of 4.3k claims, labelled as either true or false, collected from PolitiFact. As with the Snopes-2017 dataset, they queried an external search engine with the claim to collect related web documents. All in all, the dataset contains 29k articles from 336 different sources.

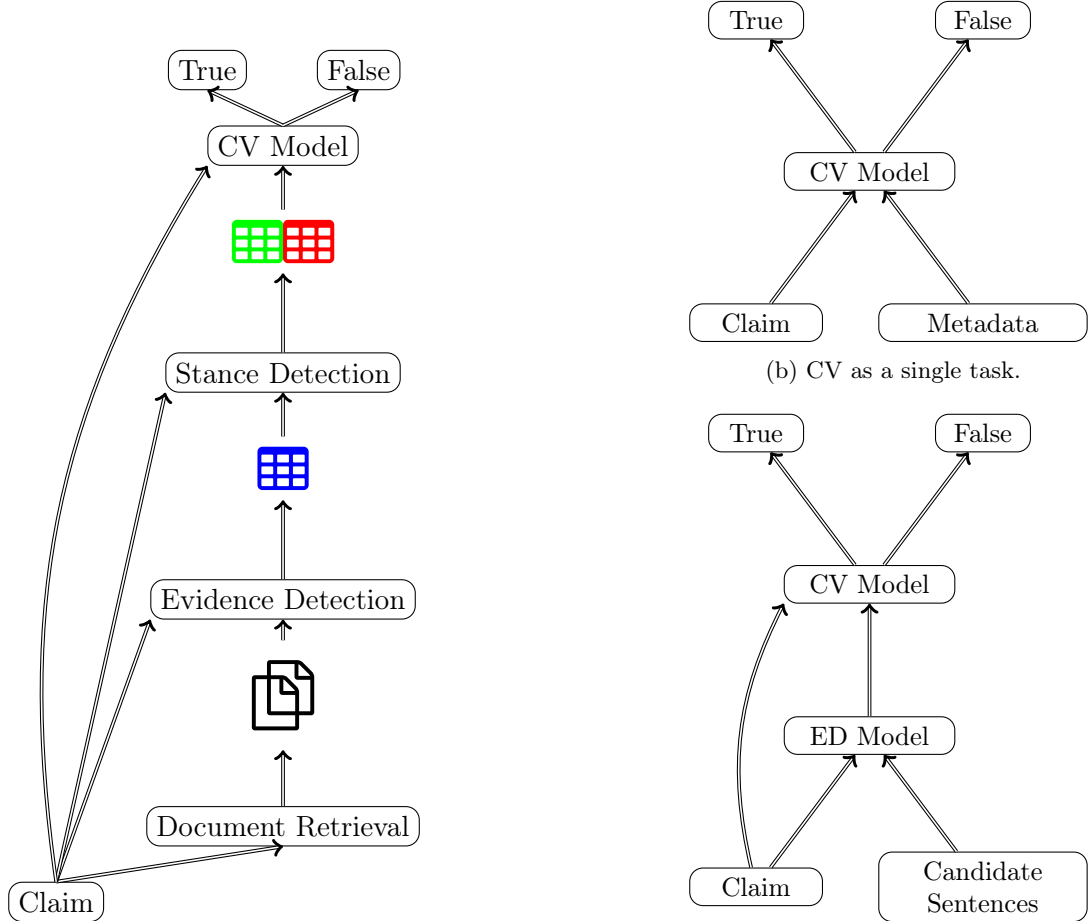
FEVER Thorne et al. (2018a) created a dataset containing 185k claims extracted and generated from Wikipedia, which are labelled as *supported*, *contradicted*, and *not enough info*. The corresponding shared task by Thorne et al. (2018b) then lists a sequence of sub-tasks and evaluation metrics. One distinction from many other CV datasets is that to sufficiently evaluate a claim, it may be necessary to extract multiple pieces of evidence from multiple source articles. For example, the claim “*The Rodney King riots took place in the most populous county in the USA*”(Thorne et al., 2018b) requires evidence from two different Wikipedia articles; one stating that the Rodney King riots took place in Los Angeles, and a second stating that Los Angeles is the largest county in the country.

Snopes-2019 This dataset presented by Hanselowski et al. (2019), contains data regarding four different tasks, namely document retrieval, document stance detection, ED, and CV. They extracted 6.4k claims from Snopes, which combined with the verdict on the claim, also contains 16k text snippets that form the evidence to support the verdict. Of

these evidence text snippets, they found 8k in the 14k source documents. They then used crowdworkers to label the directly relevant sentences in these evidence snippets, resulting in 8k sets of evidence; one set of evidence is a collection of directly relevant sentences in the evidence snippets.

3.2.2 Methods

There are three general kinds of methods used for CV: *single task*, *pipeline*, and *joint* methods. Figure 3.2 shows the different approaches to CV.



(a) CV through a pipeline. Each component in the pipeline is trained separately and the entire pipeline is only assembled for prediction.

(c) CV as a joint approach. The entire model is trained together with one set of data.

Figure 3.2: The different approaches towards CV. b illustrates a model with task exactly one task, whereas c shows a model that performs ED and CV jointly while a shows a pipeline of different pre-trained models.

Although many approaches intend to address the entirety of Claim Verification, others focus on specific tasks, such as the verification or *TE* component, which determines whether a claim is supported well enough by evidence, or detecting the stance of a document regarding the claim. Alhindi, Petridis, and Muresan (2018) extended the LIAR dataset by adding the reasons the initial human judges provided to support the automatic determination regarding which claim is true and which is false. Ferreira and Vlachos (2016) used a logistic regression with different features for the claim, the headline, and their alignment

which was calculated similarly to Rus and Lintean (2012) to predict the stance of an article regarding the claim on the Emergent dataset.

Pipeline Methods

A common approach to address determining the veracity of a claim is to use a pipeline of multiple sub-tasks. This is especially important when the collection of documents containing potential evidence is large. Given that document retrieval is a mature area of research in information retrieval, it is often used as a baseline or even without many modifications as the first step in the pipeline. Following the retrieval component, there are generally two different approaches. First, the retrieval component is followed by an ED method, which either already detects the stance of a piece of evidence or uses a separate stance detection model as a down-stream task. This means the ED component can either do a binary classification of *evidence* or *no evidence* of a sentence or document or perform a ternary classification of *supporting evidence*, *contradicting evidence*, or *no evidence*. As a final component, there is generally a TE model that predicts the veracity of a claim. Another approach follows the retrieval component with an ED model, an ED stance-detection model, and finally, a TE model (Hanselowski and Gurevych, 2017).

The former approach is followed by Hanselowski et al. (2018), who used an entity linking based retrieval of Wikipedia articles and heuristics to retrieve candidate documents. For ED, they used an Enhanced LSTM (Q. Chen et al., 2017) that they trained for ranking evidence with a pairwise hinge loss between a piece of evidence and a randomly selected sentence. They then used a hierarchy of networks with ESIMs as encoders and a combination of attention and pooling to determine the claim’s veracity.

The latter approach is followed by Hanselowski et al. (2019), who compared different methods, such as ones based on feature engineering with neural networks of different kinds, at each of the three steps following the document retrieval. They found that feature based approaches worked best for the document stance detection and that a bilstm and TF-IDF based ED reached the highest recall@5 and precision@5, respectively. For claim veracity, the best performing method was an Multi-Layer Perceptron (MLP) using embeddings generated by BERT, with the previously described extended ESIM reaching the highest macro-recall in the FEVER shared task (Thorne et al., 2018b).

Joint Methods

Others use the ED part of the pipeline to improve the CV component. In DeClarE, Popat et al. (2018) used a combination of the claim and the tokens of a document to calculate the claim attention that allows them to find the parts of a document discussing the claim. They then used the attention-weighted document, together with *claim source embeddings* and *article source embeddings*, to calculate the credibility of the article regarding the claim. Finally, they then aggregated the credibility of each article to obtain an overall credibility score for the claim.

TwoWingOS is another end-to-end ED and CV approach presented by Yin and Roth (2018). They jointly optimised the probability of a candidate sentence being a piece of evidence, and then the negative log likelihood of the prediction whether a claim is true or false and the ground truth. They used the FEVER dataset showing that it outperforms the originally presented baseline.

Ma et al. (2019) modelled ED and CV by jointly fine-tuning pre-trained models for *coherence* and TE. A sentence *coheres* with a claim if they are both addressing the same topic, hence allowing them to reduce the number of candidate sentences that might provide evidence. They calculated the coherence attention and used the TE model to calculate the

entailment attention allowing them to weigh the candidate sentences and determine the claims’ veracity on the Snopes-2017, PolitiFact-2018, and FEVER datasets.

JELTA is a probabilistic joint model of direct and indirect assertions of claims to determine the source of a claim’s trustworthiness and the claim’s veracity, presented by Zhang, Ives, and Roth (2019). Direct assertions are statements in a source that directly support or contradict a claim, whereas indirect assertions provide evidence that enables a reader to reason about the validity of a claim. The authors used a TE model as a noisy ED method and outperformed methods that only consider direct assertions, indirect assertions, or the source’s trustworthiness.

3.2.3 Relation to Research in History and Sociology

Similar to the ED, the relevance of datasets created and methods developed for CV varies between the different approaches.

Corpora

Because our goal is to support researchers in history and sociology in finding evidence and speed up their hypothesis validation, we cannot use any of the pure CV dataset, which aim at reasoning whether a claim is true or false, based on the claim alone; i.e. the PolitiFact-2014, and LIAR datasets. The stance detection datasets, namely Emergent, Snopes-2017, and PolitiFact-2018, can be relevant for EL because they contain relevance information. However, the Snopes-2017 and PolitiFact-2018 datasets both contain only documents and no evidence annotations. Although the Emergent dataset contains stance labels for shorter text snippets, rather than documents, it only consists of related snippets.

Both the FEVER and Snopes-2019 datasets contain claims as well as pieces of evidence which enable reasoning about the claim. This means that these datasets are more relevant to our work. However, their focus is clearly on the automatic CV or reasoning part, which may not be possible in research, because it misses the required contextualisation. Neither dataset contains claims that are research hypotheses, and they do not consider the existence of different schools of thought. This means that they do not contain any claim that is grounded in a particular school of thought. Therefore, the datasets can at best be used for pre-training ED and EL models.

Methods

The single task methods, especially the ones that reason about the validity of a claim, are not beneficial for our purposes, because they do not have any world knowledge and also do not rely on any kind of evidence. The pipeline approaches contain parts that can be beneficial in supporting researchers in finding evidence and validating hypotheses. However, the lack of contextualisation provided by the entailment models makes them less useful for research purposes. For instance, there is a pipeline approach presented by Hanselowski and Gurevych (2017), which consists of *document retrieval*, *document stance detection*, ED, and CV. From this pipeline, the document retrieval, stance detection, and ED components are also relevant in supporting researchers in finding evidence and validating hypotheses.

Similarly, joint methods that model ED and CV are also relevant to our research because we could leverage shared parameters to improve the performance of the ED model. However, neither approach considers the differences between individual users, and as with ED, many approaches depend on complex pre-processing pipelines or feature engineering.

Hence, we focus on adapting methods that are less dependent on language-specific pre-processing components.

3.3 Argument Mining

Argument Mining (AM) is an established field within NLP. It is concerned with the detection and analysis of argumentative content, e.g. persuasive essays or political discourse, as well as the argumentative structure or *scheme* itself (e.g. Visser et al., 2020). An argument is generally a single or sequence of statements that are logically connected for the purpose of persuading someone else. A common structure of an argument is a single *claim* and supporting or contradicting *premises*. In the claim-premise scheme, the claim is a general statement that is contested. A premise then supports the claim, which can take different forms. In research, a hypothesis and piece of evidence together form an argument in the same way as a general claim and supporting *premise* or *warrant*.

Argument Mining can generally be separated into two approaches. First, for a *discourse analysis*, all arguments made in a text are extracted. The other approach focusses on how an argument can be used by someone other than the original author, which is often called an *information seeking* perspective (Habernal, Eckle-Kohler, and Gurevych, 2014). Other research in AM investigates the aggregation of arguments and their persuasiveness (e.g. Habernal and Gurevych, 2016a).

Discourse Analysis

A large part of research in argument mining follows the discourse analysis approach, in which a specific genre of text, such as persuasive essays or online comments on public participation, are annotated with the intention of following the entire argumentation. A widely used corpus is the *UKP-Essay-corpus*, which was created by Stab and Gurevych (2014) and extended in Stab and Gurevych (2017). It consists of 401 persuasive essays written by language learners and annotated for a discourse analysis by an annotation scheme with three types: the *major claim*, *claim*, and *premise*. This means the annotation of an essay started with the *major claim*, which is the primary claim or conclusion of the essay. The next step was to find other *claims* in the essay that directly support or contradict the major claim. As the last step, the annotator labelled *premises* which support or attack a claim or another premise. A similar scheme was used by Liebeck, Esau, and Conrad (2016), who annotated comments in online participation forums. As suggested by Habernal, Eckle-Kohler, and Gurevych (2014), the methods do not generalise well across different domains (Daxenberger et al., 2017). Combining the source and target domain data or using multi-task learning improves their generalisation (Schulz et al., 2018). Frau et al. (2019) also used the essay corpus to investigate different attention mechanisms for AM. However, this is not the only argumentation scheme used in AM (see Peldszus and Stede, 2013).

Wachsmuth et al. (2017) presented args.me⁶, an argument search engine which uses an argumentation scheme similar to the claim/premise one in building a general-purpose argument search engine. They built an AM model by extracting arguments from debate portals (Ajjour et al., 2019) and conducted the argument detection in the pre-processing step. After retrieving the arguments, they used an argument-specific ranking system (Wachsmuth, Stein, and Ajjour, 2017).

⁶<https://www.args.me/index.html>

Information Seeking

In contrast to the discourse analysis approach, the information seeking perspective does not consider all arguments made within a text. Instead, given the interests of a particular user (Habernal, Eckle-Köhler, and Gurevych, 2014), a large part of the discourse can be irrelevant. Suppose a user is searching for arguments related to nuclear waste management. They might find several general debates around nuclear energy in which nuclear waste management is discussed. These debates also contain arguments around the risk of a nuclear fallout, which the user is not interested in. Therefore, the relevance of the individual argument to the user is important. Furthermore, using this perspective also means relying on many, possibly quite different sources of arguments, meaning that the AM must generalise across domains and potentially across argumentation schemes.

An example of a dataset based on the information seeking perspective is the *UKP Sentential AM* (Stab et al., 2018b). It consists of 8 controversial topics and about 25k sentences that can be supporting or attacking arguments or no arguments at all. It was created by first querying a search engine for documents relevant to a controversial topic and then annotated by ten crowdworkers to label the relation between the controversial topic and sentences in the relevant documents. It was subsequently used by Reimers et al. (2019) to show that BERT outperforms previous AM methods with a large margin; its annotation method forms the basis for the argument search engine ArgumenText (Stab et al., 2018a). Args.me (Wachsmuth et al., 2017) also uses the information seeking perspective by querying for and ranking the relevant arguments from heterogeneous sources; although only in down-stream tasks after mining all the arguments from documents.

Argument Aggregation and Argument Persuasiveness

When mining arguments from many different documents, one runs into the problem that many arguments are redundant. For instance, in different documents on the topic of *nuclear energy*, the argument “*Nuclear energy releases less carbon dioxide than coal and is therefore more environmentally friendly*” can occur in documents on nuclear energy as well as climate change. However, a user might not need to see every occurrence of each argument. Instead, an aggregated view can be informative in two ways. First it reduces the clutter, so the user can see the relevant information, and second, the number of arguments in each aggregation can be used to inform about how often an argument is used. There have been several approaches for argument aggregation. A naïve approach uses semantic textual similarity, a task in which two statements are rated regarding whether or not they mean the same thing. However, Boltužić and Šnajder (2015) found that using it in combination with agglomerative hierarchical clustering does not result in a high performance because of a lack of background knowledge; hence, the authors suggested the possibility of arguments to touch different aspects at the same time and suggested using an argument-specific similarity. Misra, Ecker, and Walker (2016) approached this by using *argument facet similarity*. An argument facet is a category in which different arguments can be made, such as morality or constitutionality. They then used crowdworkers to label the argument facet similarity of pairs of arguments. Reimers et al. (2019) used a similar approach but with three labels to create a second aspect-similarity-based dataset; they found that similarity prediction from a pre-trained BERT (Devlin et al., 2018) model work considerably better than other features for agglomerative hierarchical clustering.

When selecting an argument within a discussion or general debate, its persuasiveness is key. Using an argument that is not very persuasive would not lead to the desired result in influencing or changing another person’s point of view. Andrews et al. (2006) proposed to use persuasive arguments to improve chatbots for healthcare advice. Datasets are often

created in a preference based setup. This means that a crowdworker is presented with two arguments and is supposed to select the more persuasive one. Such an approach was used by Habernal and Gurevych (2016b) to create a dataset of 35 topics and 16k pairs of arguments. They also found that the resulting argumentation showed properties of “a strict total order”. Gretz et al. (2019) used crowdworkers to first generate two arguments, one supporting and one contradicting a given topic. They then used a different group of crowdworkers to label the arguments as whether they would suggest using it in a speech on the topic. However, Lukin et al. (2017) found that argument persuasiveness is not perceived by everyone as the same. Instead, they found a connection between the persuasiveness of an argument and the personality type.

Relation to Research in History and Sociology

Given that a hypothesis evidence pair is a special kind of argument, the datasets and methods used for AM can also be relevant to the goals of this thesis. The information seeking perspective, in particular, is relevant to the detection of pieces of evidence and linking them to hypotheses. Finding evidence is similar to finding arguments in the information seeking perspective; the sources are very diverse and might contain evidence relevant to other hypotheses but not the ones of a particular user. Furthermore, AM with the information seeking perspective can be separated into two tasks: *argument detection* and *argument topic linking*. Therefore we use the UKP Sentential AM dataset in future experiments.

Aggregating arguments as pieces of evidence can also be of great benefit in supporting researchers in history and sociology. For instance, aggregating arguments over time can be used to show how a discourse has changed over time or how a particular argument is adopted by a society. Argument persuasiveness might be useful in supporting a researcher in writing more convincing research articles. However, both approaches are outside the scope of this thesis.

3.4 Other Related Research

This thesis’ focus on supporting researchers, especially in the humanities and social sciences, means that it is closely related to other fields of research. It is related to *interactive machine learning* because we are aiming at directly supporting real users. This means we will most likely have to learn interactively from the user. It is also related to the field of *digital humanities*, which combines computational and humanities perspectives and methodologies. However, even though our research is primarily in ED and NLP, it still makes contributions to other fields.

Interactive Machine Learning

Although NLP is supposed to support users, little research include users in their actual design. For instance, a lot of research in AM or ED has not considered a user in their evaluation or any other part, even though the purpose of either task is oriented towards an actual human user. Other fields of research, such as summarisation, use metrics that correlate with human ratings like ROUGE (C.-Y. Lin, 2004) or include simulations of users due to the high cost involved in human research participants. However, involving humans in the evaluation or the actual research design is common in other fields of work. In robotics, for instance, learning from demonstrations depends on a human to demonstrate how to solve a task (Schaal, 1996; Atkeson and Schaal, 1997); here, human insight can be

used to guide a learner (Thomaz and Breazeal, 2008) or to reduce the necessary amount of fine-tuning of hyper-parameters (Stahlhut et al., 2016). User interaction is also vitally important in *information retrieval*, especially in learning how to rank (Zoghi et al., 2017) and to handle concept drift (Kangasrääsiö et al., 2016). In NLP, many approaches either focus on the learner and its task, i.e. have a *system focus*; few focus on the user of the system, i.e. have a *user focus*.

System focus In active learning, the learner is in control and queries an oracle, either a human or simulated one, with the goal of learning the fastest. Oftentimes, the learner has a limited number of queries, and much research went into finding the best method to select the next sample to query the oracle. The selection of a sample is often based on the certainty of a model or on an ensemble of models (Lughofer, 2011). If the ensemble shows a large disagreement on a particular sample, it is presented to the oracle. In the APRIL framework, Gao, Meyer, and Gurevych (2018) used a combination of active learning and preference learning to build a user-specific summarisation reward function which they applied to multi-document summarisation. They combined these two approaches by using preference learning in letting a user decide which of two summaries they preferred before ranking the results so that a reward function can be learned from it. The active learning component is used to select the candidate pair to present to the user. In this way, they avoid querying the user needlessly with uninformative candidates. Kasai et al. (2019) combined active learning with transfer learning to reduce the required number of queries to the oracle.

User focus Focussing on the user means supporting them in reaching their goal and keeping them in control. Although the machine decides which sample to query the user with in many active learning approaches, the user decides independently from the machine what to annotate next. This can be seen in the INCEpTION platform (Klie et al., 2018) and in supporting medical experts in evaluating student’s diagnostic reasoning (Schulz et al., 2019). Another example is supporting users in creating an entity graph in the medical domain (Yimam et al., 2017) or multi-document summarisation (P.V.S. and Meyer, 2017).

Combining User and System focus With AlpacaTag, B. Y. Lin et al. (2019) used two different components in enabling faster annotation of sequence tagging tasks from crowdworkers. First, they supported the crowd-worker in using the already annotated samples to train a commonly used sequence tagging architecture, i.e. a bilstm with a conditional random field, using its predictions to make suggestions to the crowdworker. Allowing the crowdworker to accept or reject the suggestion supports them in creating the sequence tagging labels. On the system side, on the other hand, they used an active learning based strategy to select the samples for annotation for the crowdworkers. Lee, Meyer, and Gurevych (2020) combined the system’s need of receiving the most informative sample with the user’s need of receiving an adequate sample in the creation of c-tests. In a c-test, a language learner is tasked with filling the gaps in a text. The gaps can be individual characters or sequences of characters at the beginning or end of a word. When automatically generating c-tests from texts, the system’s focus is on requesting the most informative sample to reduce the number of queries. However, this can lead to inadequate c-tests for a language learner; too easy and the learner becomes bored, too hard and the learner becomes frustrated. Lee, Meyer, and Gurevych (2020) used two sampling functions, namely *uncertainty sampling* to model the system’s focus and *appropriateness sampling* to model the user’s focus. They then combined these sampling strategies to define the actual sampling strategy.

Digital Humanities and Using NLP in Humanities Research

The growing field of *digital humanities* is a collaborative and transdisciplinary field of research combining the humanities with computational research, but is it not just the application of computational methods for research in the humanities. “Rather, Digital Humanities is defined by the opportunities and challenges that arise from the conjunction of the term digital with the term humanities to form a new collective singular” (Burdick, 2012, p. 122). Examples that combine research in the humanities, as well as NLP are:

Garg et al. (2018) used a text corpus spanning 100 years to quantify gender and ethnic stereotypes. They used, among others, word embeddings calculated on Google Books / Corpus of American English for each decade in the 20th century and determined the distance between occupations and male or female description words, such as he and female. They also studied the bias towards specific ethnicities by last names associated with a particular ethnicity instead of gender pronouns. This line of research is important for both the humanities and NLP for two reasons. First, it quantifies existing bias in historical texts and their development, pointing towards specific periods, such as the *second wave of feminism* in the 1960th. Second, it demonstrates that word embeddings inherit biases present in their training corpora and that embeddings trained on historical texts may not be applicable for modern NLP problems. Furthermore, it also shows how much the meaning of individual words changes over time, which creates additional challenges when processing text from different time periods in a single NLP pipeline. Detecting and reducing biases has now become an established problem within NLP (Prost, Thain, and Bolukbasi, 2019; Sweeney and Najafian, 2019; Zhao et al., 2019).

Rutherford and Thanyawong (2019) used a maximum likelihood search system and text of known literary era of *old thai prose* to contribute both to NLP and literary research. In NLP, they contributed with their method of building an interpretable model to identify in which literary era a piece of text has been authored. In literary research of Old Thai Prose, they presented the first classification of text into literary epochs.

3.5 Chapter Conclusion

In this chapter, we analysed how current research in NLP approaches the tasks of ED, EL, CV, and AM. We found that although there are several datasets for ED and CV, neither directly aims at supporting researchers in history and sociology. Still, some datasets are more applicable than others. Whereas some datasets are for both, ED and CV are annotated on a discourse level, meaning that all pieces of evidence in a document are annotated, while others follow the information seeking perspective in which only topic relevant pieces of evidence are annotated. We think that the information seeking perspective is more similar to research; therefore, the datasets based on this perspective are more applicable to our research aims. However, neither dataset considers the existence of different schools of thought, which means that their direct benefit might be limited; therefore they may be more useful for pre-training to address the cold-start problem.

We also analysed how the methods developed in current research can be applied to supporting researchers in history and sociology. We found that all methods developed for CV, ED, and AM use a static task definition in which the goal of the user does not change. However, especially in sociology, this is not the case. Methods, such as grounded theory depend on a fluid task definition because the goal of the user changes as they carries out their research. This is rarely addressed in NLP research, with *concept shift* being the most similar.

In analysing contemporary research in IML, we found that many approaches require a

user but focus on the goal of the machine. For instance, this can be seen in letting the user select which of two summaries they prefer. These preferences are then used to train a reward function that is used by an extractive summarisation model. Although the user receives a personalised summarisation model, the goal of the interaction is only to create training data, not to directly support the user. However, other approaches do focus on the user's goals, but presume that the task can be solved independently from the individual user.

We conclude from our analyses that to support researchers, we need to create user-specific datasets of actual historians and sociologists so that we can develop and evaluate the methods for ED and EL. We also need to investigate how researchers in these fields carry out actual research, because research practice oftentimes deviates from research concepts.

Chapter 4

Human Strategies for Evidence Detection

The standard approach for tasks, such as ED and EL, is to use large datasets to train models and then use their predictions to support the users. In this case, it would mean that the models have been trained beforehand and remain static while the user benefits from them. However, in chapter 2, we found that not only are there different schools of thought, but some research methodologies are extremely fluid in nature. This begs the question of whether we can use a static model or have to use one that can dynamically adapt to its users. Furthermore, the individual behaviour might also influence the requirements for a model. If a user strictly follows the Hypothetico-Deductive (HD) approach and defines a hypothesis before searching for evidence relevant to this one hypothesis —and only this one— it would be beneficial to include the hypothesis in the ED model. For a user who has multiple hypotheses in mind, on the other hand, it would not be possible to immediately link pieces of evidence to a hypothesis; in this case, an ED model would not benefit from knowing about a hypothesis.

Therefore, we conclude, that before we can support researchers in finding evidence so that they can evaluate their hypotheses faster; we must first understand how they approach these tasks. How do they go about finding evidence, and how do they use this evidence to evaluate their hypotheses? Do they read first and then link pieces of evidence to hypotheses, or do they follow a different approach? Furthermore, it is unknown whether the researcher’s goals remain static or change while collecting evidence and validating their hypotheses. Finally, we need to ascertain how well they agree on what constitutes evidence to investigate the possibility of building a gold-standard dataset of ED for researchers in the humanities and political and social sciences. To reach this understanding and thereby answer our first research question (RQ 1: How do researchers in the humanities and social sciences validate their hypotheses?), we need to answer the following sub-questions:

- ① Do researchers distinguish between finding evidence and linking it to hypotheses?
- ② Do they read a document multiple times or just once?
- ③ Do they work on multiple hypotheses in parallel or one at-a-time?
- ④ Do researchers revise their hypotheses over time, thus developing them, or are the hypotheses defined once in the beginning and then remain static?
- ⑤ How well do they agree on the evidence?

To answer the first three sub-questions, we conducted two user studies with students in the humanities and social sciences in a teaching-seminar on environmental catastrophes. Since the seminar was conducted in German, we selected speeches from the German parliament as textual sources. To log the user behaviour and collect their annotations, we developed a tool named EDoHa (Evidence Detection fOr Hypothesis vAlidation) that allow a user to annotate evidential sentences in documents and link them to self-defined hypotheses. While the user is working, EDoHa creates a log of relevant activities that allows us to analyse the user’s behaviour. To answer the fourth sub-question, we investigated whether or not researchers changed their hypotheses in EDoHa’s log files and compared the results against self-reported answers to the following question: “Did you revise your hypotheses during the study?”. We also collected the annotations the users created and the user-specific links between self-defined hypotheses and pieces of evidence to compile two datasets; one on the topic of *nuclear energy* and one on the topic of *forest dieback*.¹ To answer the fourth sub-question, we calculated the inter-annotator agreement on the first dataset for similar hypotheses.

4.1 User Study

We conducted two user studies in the context of a teaching-seminar on the topic of *environmental catastrophes in the second half of the 20th century*. The participants of the seminar were students in their mid-bachelor’s to early master’s programme with a study subject of history or political or social sciences, as well as students in dual programmes. During each normal session of this weekly seminar, a group of students gave a presentation on one environmental catastrophe, its perception, or how it affected the people afterwards. These presentations can have different topics, such as the nuclear disaster in Chernobyl or the North Sea flood of 1962 in Hamburg. During two sessions, the first one on the topic of *nuclear energy* and the second one on the topic of *forest dieback*, we asked the participants to come to our computer lab to conduct the user study.

4.1.1 Data Sources

For each study, we selected two sessions from the German parliament on a topic related to the topic of the user study. We selected speeches from the German parliament for three reasons. First, the speeches are in the public domain, allowing us to publish the data without having to consider the copyright of the individual speakers. Second, the political debates span the second half of the 20th century. Third, environmental catastrophes are a recurring topic in the parliamentary debates.

For the first selection of the data sources, we decided to search the *Parlamentsdokumentation* of the German parliament² for related terms, such as *Tschernobyl* and *Fukushima*. We then looked through the results and selected two sessions that were separated by several years. On the topic of nuclear energy, we selected a session from June 6th 1986, and June 30th 2011. The session from 1986 was shortly after the Chernobyl disaster and had the topic *health and ecological consequences of the reactor accident in Chernobyl for the citizens of the Federal Republic of Germany*. The second session took place shortly after the Fukushima catastrophe with the topic *nuclear phase-out and the 25th anniversary of*

¹Our datasets are in German, because the seminar was conducted in the German language. The term forest dieback refers to a loss of large amounts of trees in the 1980s. The common German term is *Waldsterben*.

²<https://pdok.bundestag.de/>

the *Chernobyl disaster*. We selected the first speech of a member of each political party present in the German parliament at the time.

On the topic of forest dieback, we downloaded all minutes of the German parliament, extracted the plain text from the PDF documents via Apache PDFBox³, and segmented the speeches into sentences. We then searched for the term *Waldsterben* in all documents, selecting the session with the highest count of the term and last session in which the term occurred at least five times. This means, we selected the minutes from May 20th 1983, and March 15th 2001. The first session had the topic *emergency programme against forest dieback* in combination with other topics related to the forest dieback. The second session was on the *forest condition study of the German parliament* from 1999. From the first session, we selected the first speech of a member of each political party present in the German parliament at the time. From the last session, we selected the speech of the German Secretary of State as the representative of the government and the speeches from the CDU/CSU and PDS as the representatives of the opposition. Table 4.1 shows the date, speakers, and their political affiliation for each topic.

Topic	Date	Speaker	Party	Sentences
Nuclear Energy	June 6 th 1986	Dr. Alfred Dregger	CDU/CSU	43
	June 6 th 1986	Norbert Eimer	FDP	39
	June 6 th 1986	Volker Hauff	SPD	38
	June 6 th 1986	Hannegret Hönes	Die Grünen	41
	June 30 th 2011	Eva Bulling-Schröter	Die Linke	61
	June 30 th 2011	Marco Bülow	SPD	69
	June 30 th 2011	Marie-Luise Dött	CDU/CSU	69
	June 30 th 2011	Michael Kauch	FDP	50
	June 30 th 2011	Jürgen Trittin	Bündnis 90/Die Grünen	72
	May 20 th 1983	Gerhart Baum	FDP	169
Forest Dieback	May 20 th 1983	Wolfgang Ehmke	Die Grünen	130
	May 20 th 1983	Volker Hauff	SPD	137
	May 20 th 1983	Paul Laufs	CDU/CSU	72
	March 15 th 2001	Matthias Berninger	Parlamentarischer Staatssekretär	89
	March 15 th 2001	Eva Bulling-Schröter	PDS	36
	March 15 th 2001	Albert Deß	CDU/CSU	61

Table 4.1: The specific speeches used as the data source in our user studies.

4.1.2 Study Setup

After the participants arrived at our lab, we asked them to fill out a consent form that detailed the purpose of the study and that any data collection would be done anonymously. Appendix A shows the complete consent form we used. We then gave a short introduction to EDoHa before handing out the description for their task. The paper sheet also contained the credentials for the accounts we created beforehand to anonymise the participants. Afterwards, we answered all questions the participants had regarding their task or about EDoHa. Then, the participants had about an hour to complete their task. Afterwards, they filled out a questionnaire, followed by a discussion of their findings regarding the task given to them. Appendix A shows the task description and the questionnaire for the first user study.

³<https://pdfbox.apache.org/>

First study In the first user study, we asked the participants to compare the political discourse after the Fukushima catastrophe with the one after the Chernobyl disaster. We also provided three guiding questions, namely:

1. Which topics are discussed controversially only after a single event?
2. Which topics are discussed controversially after both events?
3. What connections are established between the events?

To keep the study realistic and not influence the users, we did not narrow down the task any further. This means that we did not provide them with sample hypotheses to start with and tried not to suggest any approach they were supposed to follow.

Changes in the setup after the first study During our first study, we noticed that most users did not formulate complete hypotheses, but instead formulated questions and most often topics or concepts. We therefore decided to remind the students of the definition of a hypothesis, i.e. a hypothesis is a testable claim that can be supported and contradicted: we asked them to fully formulate all hypotheses. We also provided the sample hypothesis “*Adherence to nuclear energy prevents a rational transformation and leads in the long term to economic and environmental damage*”. We included no other additional guidance regarding how they should approach their task.

Second study In the second user study, we asked the participants to analyse the political discourse around the topic of the *forest dieback*. We again provided three guiding questions:

1. What is the reason / are the reasons for the forest dieback?
2. How are the reasons different at a later point in time?
3. How did the perception of the forest dieback change over time?

Table 4.2 shows the number of participants, i.e. humans participating in our study, and the users logged into EDoHa. The numbers in the first study are not identical, because we had two groups of two participants each, and we treat these groups as users.

Dataset	Participants	Users
Nuclear Energy	18	16
Forest Dieback	13	13

Table 4.2: Number of participants and users in both user studies.

4.2 EDoHa

To study the user’s approach in validating their hypotheses, we first need to analyse it. This means that we need the participants of our study to search through documents, mark pieces of evidence, and link them to self-defined hypotheses. We then need to analyse their behaviour to understand how they approached these tasks, particularly ED and EL. We therefore developed a tool that would enable the users to mark evidence and link it to self-defined hypotheses. Our tool EDoHa is based on the annotation platform INCEpTION (Klie et al., 2018). INCEpTION is a Java-based annotation tool that allows the user to create annotations in different granularities, such as individual tokens or spans of tokens

within and across sentence boundaries. It also contains components to merge these annotations made by different users and make suggestions to users. We extended the platform by creating a new user interface for evidence annotations and the ability to log the user's activities to enable a behaviour analysis.

4.2.1 User Interface

Our extensions to the user interface consist of a document list and two different views: the *document view* and *evidence linking view*. The document list shows all documents in the current project with the one that the user is currently annotating as highlighted with a green font (①). For each document, the list shows the topic, the name of the speaker, the political affiliation, and the session ID of the specific day. The icons at the bottom right of each document indicate whether the user has opened the document and if there is the existence of manually annotated pieces of evidence. The user can unselect the current document by clicking on the **Clear Selection** button (②) at the top right of the document list. The *document view* allows the user to label sentences as evidence, and the *evidence linking view* allows for linking these pieces of evidence to self-defined hypotheses.

Document View

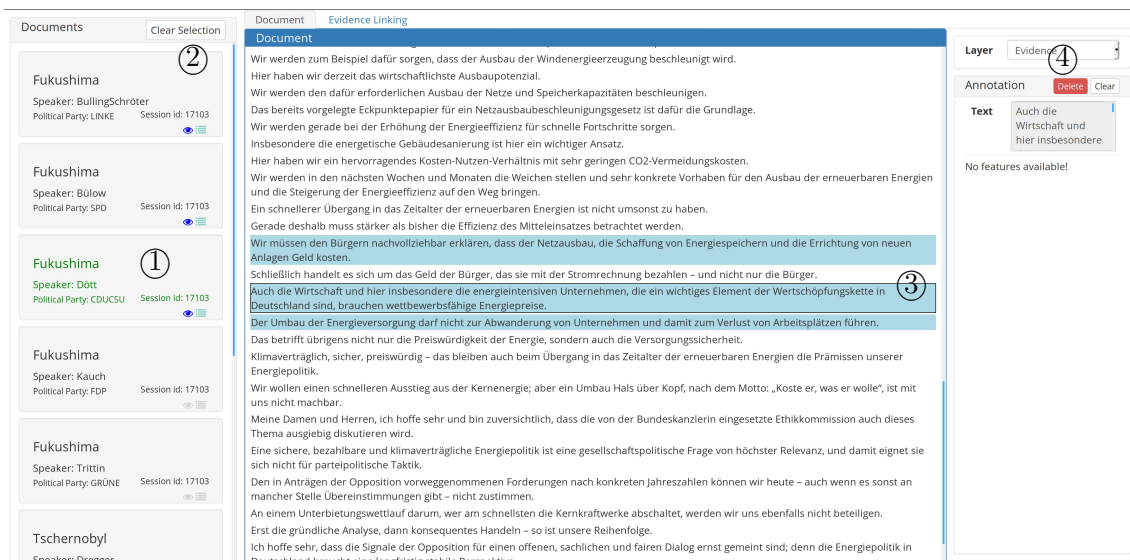


Figure 4.1: Document view of EDOHa. The numbers refer to the individual components described in the main text.

The document view (figure 4.1) shows each sentence of the document as an individual line of text. The user can then label a sentence as a piece of evidence by clicking on it, which is then signified by a blue background colour (③). A user can remove their label by clicking on a labelled sentence and then clicking on the **delete** button in the right-hand column (④). The currently selected sentences are also highlighted with a black frame. The far right column shows the properties of the currently selected piece of evidence. During the first user study, we used an annotation editor provided by INCEpTION that is based on *brat* (Stenetorp et al., 2011), but we found that many users struggled with labelling complete sentences. To label a complete sentence, the *brat* interface requires the user to highlight the complete sentence with the cursor. However, when the user moved slightly off the current sentence, the *brat* interface highlighted additional tokens in preceding or

succeeding sentences. This led us to develop the sentence-level annotation editor we used in the second study.

Evidence Linking View

Figure 4.2 shows the *evidence linking* interface in which a user already created three hypotheses and linked multiple pieces of evidence to it. The left column is identical to the document view and shows the list of available documents. The top half shows the list of hypotheses the user has defined, and its header contains a text field in which the user can specify a new hypothesis (①); the bottom half shows the list of evidence the user has found. Once the user has created a hypothesis, it is shown as a cell below with the title of the cell containing the hypothesis (②), which the user can modify, if desired. The user can then link pieces of evidence to this hypothesis via Drag & Drop from the list of evidence at the bottom of the window (③). To avoid cluttering the user interface, we limited the length of each piece of evidence to 90 characters and show the complete sentence as a tool tip if the mouse hovers over it.

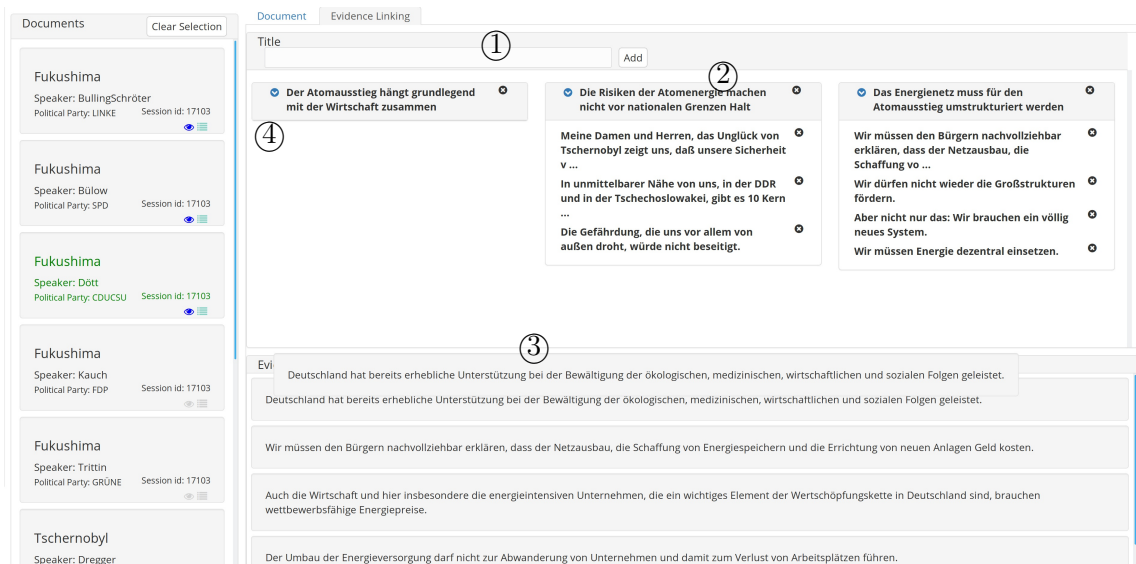


Figure 4.2: Evidence linking view of EDoHa. The numbers refer to the individual components described in the main text.

On the bottom half of the screen, we can see the list of evidence that the user has labelled in the currently selected document. By selecting a different document, the user can switch to another source of the listed evidence. The document will then be highlighted with a green font. We limited the selection of evidence to one particular document to avoid having very long lists; if the user wishes to see all the evidence from all documents, they can click on the **Clear Selection** button above the document list. If the user wants to remove a particular piece of evidence from a hypothesis, they can click on the small cross on the right-hand side of the evidence (✕). Clicking on the cross on the right-hand side of the title deletes the entire hypothesis and all links to the linked evidence. The blue down arrow folds the hypothesis up. This means that after clicking on it, only the hypothesis text is visible and the list of linked evidence is not (④). This also allows for keeping the user interface clean and clutter-free.

In early tests with expert users, we found that a method to find the source of a particular piece of evidence would be beneficial. Therefore, we implemented a highlighting mechanism

that allows the user to see which document a particular piece of evidence has originated from. If the user hovers the cursor over a piece of evidence in the list of the linked evidence of a hypothesis and presses the Alt key, EDoHa shows a dashed frame around the document, the particular piece of evidence, and all hypotheses that this piece of evidence is linked to. This also allows the user to avoid linking a piece of evidence to a hypothesis to which it is already linked. If a piece of evidence is already linked to a particular hypothesis, then the hypothesis will be highlighted and the user does not need to create another link. This is particularly beneficial if the hypothesis is folded up.

4.2.2 Data Import

We decided to perform sentence segmentation before importing the documents into EDoHa. This allowed us to reduce the computational load because of redundant sentence segmentation each time a document is opened. We created a UIMA pipeline with DKPro Script (Eckart de Castilho and Gurevych, 2014) that read the plain text source documents, segmented them into sentences, and tokenised them. The resulting UIMA XMI files can then be loaded into EDoHa. We selected Apache OpenNLP⁴ for the sentence segmentation and tokenisation because it is available in multiple languages. Using DKPro Script also allows us to conduct more complex pre-processing, if the need arises. We included the entire pre-processing script in the appendix as listing B.1, with `$DATA_FOLDER` representing the folder containing the plain text files.

4.2.3 User Activity Logging

To analyse the behaviour of the users, we logged many different activities, such as opening a document and creating or modifying a hypothesis. Table 4.3 shows which events we logged. All events are also saved with a timestamp and user who triggered the event.

4.3 User Behaviour

We conducted five analyses of the users' behaviour to understand how they developed and validated their hypotheses. First, we looked at whether or not they showed a distinction between phases of ED and EL. Second, we explored whether they focused on a single document and, once finished, if they visited it again or worked on multiple documents at the same time. Third, whether they worked on multiple hypotheses at the same time or focused on a single one. Additionally, we looked at whether there is a correlation between users who worked in phases of ED and EL and on single hypotheses. As the last part of the analysis, we looked at how the hypotheses change, thus developing over time.

4.3.1 ED Phases vs. No Phases

When analysing the timing at which each user annotated a sentence as a piece of evidence and when they linked a piece of evidence to a hypothesis, some users show a distinction into phases of evidence detection and hypothesis validation. Some users first created a number of annotations before switching tasks from ED to EL by linking several previously labelled pieces of evidence to self-defined hypotheses.

Figure 4.3 shows the number of sentences that were labelled as evidence and the number of links between hypotheses and evidential sentences over time. Each blue dot represents the act of annotating a sentence as a piece of evidence, with the y-axis showing how many

⁴<https://opennlp.apache.org/>

Event Trigger	Event Properties
Open Document	The opened document
Evidence annotation created	Source document, content, and span position + “created”
Evidence annotation deleted	Source document, content, and span position + “deleted”
Created link between hypothesis and evidence	Hypothesis text and evidence text + “created”
Deleted link between hypothesis and evidence	Hypothesis text and evidence text + “deleted”
Updated text of hypothesis	old and new hypothesis text
Created Hypothesis	Hypothesis text + “created”
Deleted Hypothesis	Hypothesis text + “deleted”
Changed tab to document view	“annotation”
Changed tab to hypothesis validation view	“hypotheses”
Clear Selection button	“clear-selection”

Table 4.3: The events logged in EDoHa to analyse user behaviour. The texts in quotes are constants.

sentences in total were annotated by this user. Each orange cross shows the creation of a link between an evidential sentence and hypothesis that a user created; again, the y-axis represents the total number of links for this particular user. The green dots show when a new hypothesis was created.

The distinction among the phases is especially apparent for User14, who basically only had one phase for each task. User7 and User0 also showed clear signs of different phases, in which pieces of evidence were labelled and then linked to the hypotheses. User15, User19, and User17 do not show such a distinction. Each time they labelled a piece of evidence, they also created at least one link between a hypothesis and piece of evidence. User19 even created more links between hypotheses and pieces of evidence than they labelled pieces of evidence. However, we can also see that the duration of each phase shortened as time passed, which is particularly apparent with User7. This can indicate that a phase-free approach is actually more natural to the users, or that they intended to reduce the cognitive load or time it takes to find a piece of evidence in the list of labelled pieces of evidence.

We also found that some users first created several annotations and subsequently deleted them, for instance, User7 or User16. Because this appears only at the beginning, we conclude that it shows users learning to use the annotation editor. Some users, User2 in particular, did not show this initial learning phase; however, this does not mean that they encountered no problem with the annotation editor, only that if they made a mistake, they did not correct it. This observation and the feedback provided by the users led us to build a new annotation editor that is easier to use.

Furthermore, only a few users showed the creation of multiple hypotheses in the beginning, before starting to search for evidence, indicating to follow an HD approach towards finding evidence. We can find evidence of this behaviour in User20, though User13 and

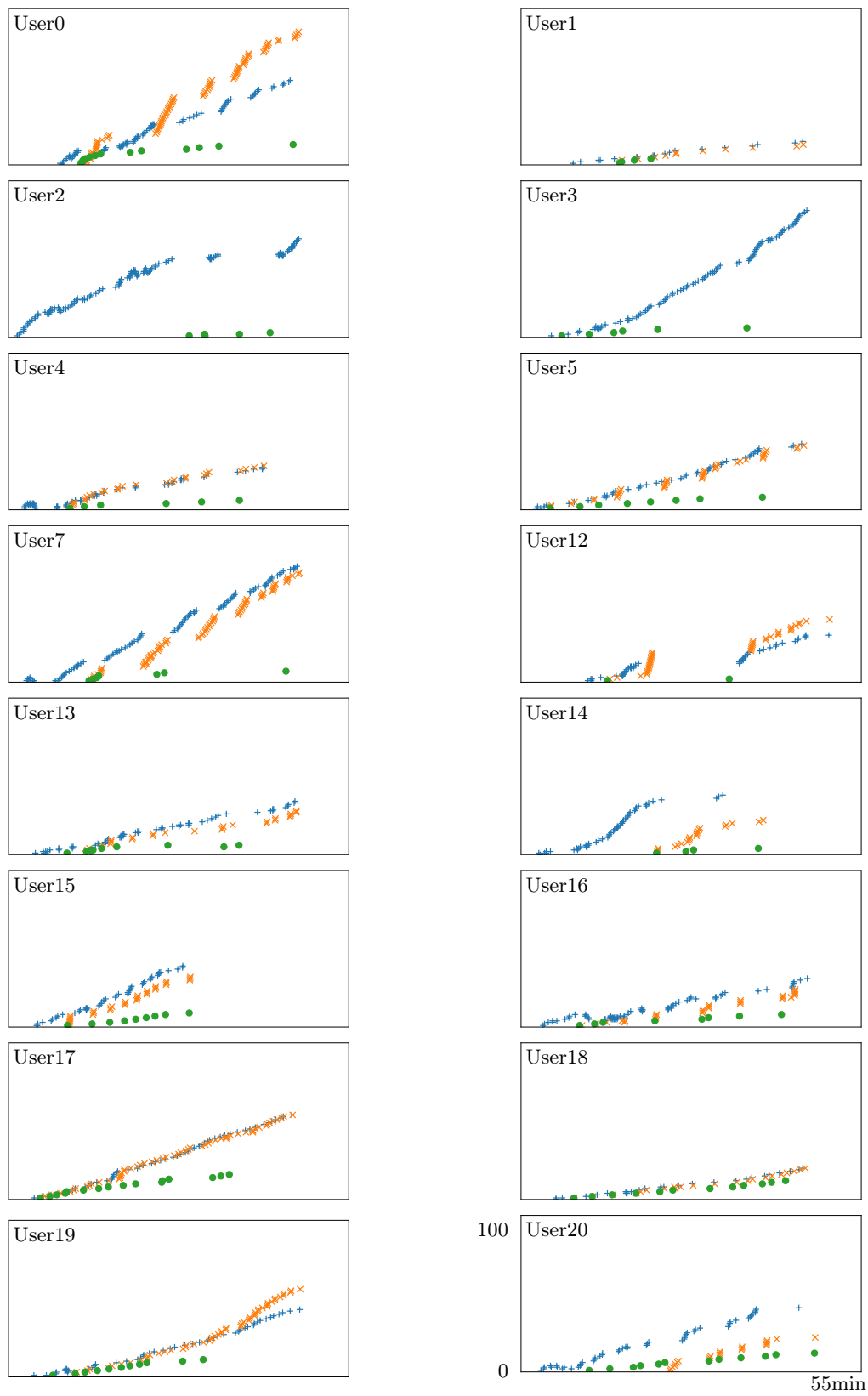


Figure 4.3: The number of evidence annotations (+), evidence/hypothesis links (x) for validation, and hypotheses (•) over time.

User19 also created at least one hypothesis in the beginning before searching for more evidence. User2 and User3 created no link between a piece of evidence and a hypothesis.

4.3.2 ED in One Document at a Time vs. Multiple Documents in Parallel

Another point of interest is whether the users worked on multiple documents at the same time or focussed on one document and, once finished with it, did not return to it. Therefore, we took all **create annotation** and **delete annotation** events and the documents in which these events occurred and plotted them over time. Figure 4.4 shows all these events and the documents in which they occurred with each document being a row in alphabetical order, starting with the first document at the top.

A large majority, in fact all users except for four, namely User4, User12, User13, and User14, never revisited a document once they were finished labelling the evidence in it. Furthermore, many users worked through the documents in alphabetical order, which is how the documents appeared in EDoHa. This is particularly apparent for users User3 and User7. User0, on the other hand, started in alphabetical order but at a later point focussed on documents that were at the bottom of the list of documents in EDoHa. We assume that this because of the documents at the top being about Fukushima and those at the bottom about Chernobyl, and the user wished to cover both events before the exercise was over.

4.3.3 Evidence Linking Approaches: A Single Hypothesis at a Time vs. Multiple Hypotheses at a Time

To analyse whether the users worked on a single or multiple hypotheses at the same time, we plotted the number of pieces of evidence linked to a particular hypothesis over time. We ignored changes of a hypothesis' text and treated it the same before and after a change. This means that once a hypothesis was created, we counted all added and removed pieces of evidence; if the text of the hypothesis was changed, we considered it the same as before.

Figure 4.5 shows a variation of the user's approaches to linking pieces of evidence to hypotheses. Some users, particularly User15 and User12, worked on only a single hypothesis at-a-time. User12 created two hypotheses, the first in the beginning and the second after about 40min. In between, the user linked labelled pieces of evidence only to the first hypothesis and once the second one was created, they linked pieces of evidence only to the second one. Other users, such as User7 and User0, worked on multiple hypotheses at the same time. Most users also created a few more hypotheses at a later point in time, indicating that they used a mixed approach of defining hypotheses and finding evidence. We compared this result with their answers in a questionnaire, regarding whether or not they first defined a hypothesis and then searched for evidence, or first collected evidence and then formulated a hypothesis later. In this comparison, we found that most users followed a mixed approach while still reporting that they actually either collected evidence first or formulated a hypothesis first.

4.3.4 Correlation between Phases and Other Behaviours

We found that some users worked in distinct phases of ED and EL, while others showed no phases, but instead linked a piece of evidence they annotated directly to one or more hypotheses. Because we also found users working on single documents or single hypotheses at-a-time, we investigated whether or not there is a correspondence between working in phases and on single documents or single hypotheses at-a-time. Table 4.4 shows whether

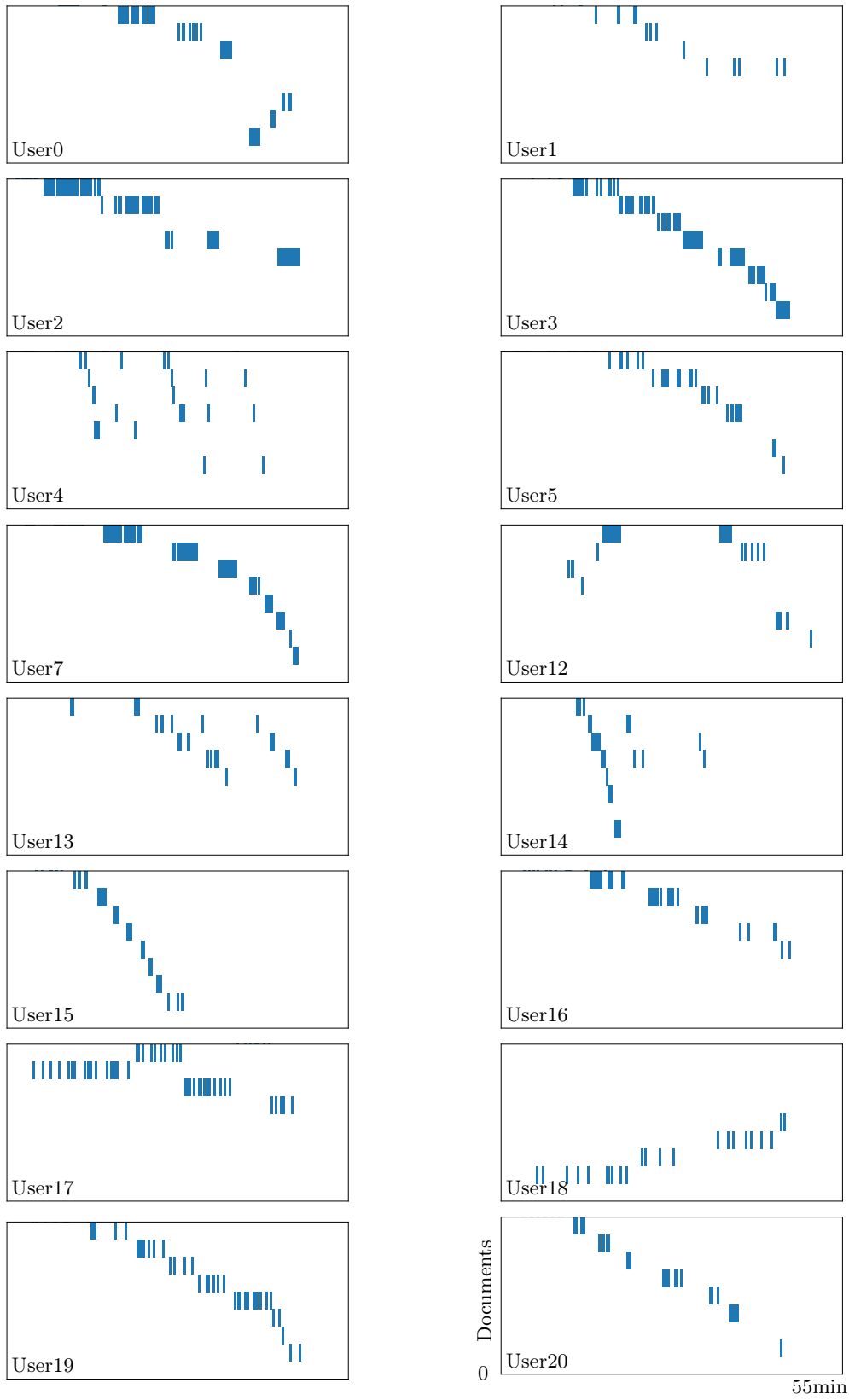


Figure 4.4: Activity of the users in documents over time. Each row represents one document sorted alphabetically from top to bottom. Each vertical bar represents an event, either the creation or deletion of an annotation, in the particular document.

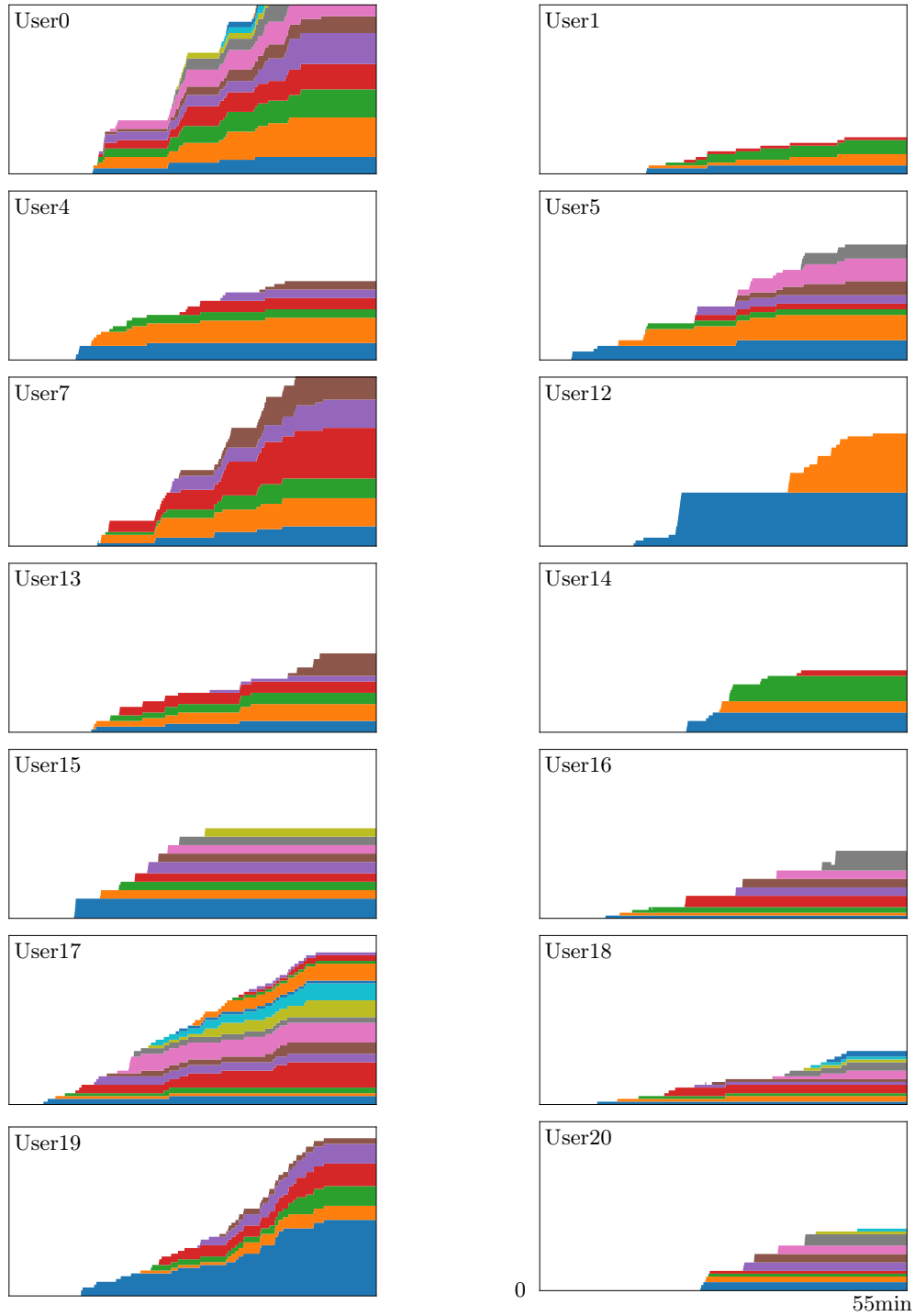


Figure 4.5: The growth of hypotheses for each user over time. Each layer represents a hypothesis, and the height represents the amount of evidence linked to it. User2 and User3 did not create any link between a piece of evidence and a hypothesis.

a user approached the task in separate ED and EL phases and whether this user kept only one hypothesis in mind.

User	Phases	Single Hypothesis	Single Document
User0	yes	no	yes
User1	no	no	yes
User2	no	no	yes
User3	no	no	yes
User4	no	no	no
User5	yes	no	yes
User7	yes	no	yes
User12	yes	yes	no
User13	yes	no	no
User14	yes	yes	no
User15	yes	yes	yes
User16	yes	yes	yes
User17	no	no	yes
User18	no	no	yes
User19	no	no	yes
User20	yes	yes	yes

Table 4.4: The behaviour of the users with ED and EL phases and whether this user worked on exactly one hypothesis or document at-a-time.

Based on the distribution of correlating behaviour, we created three groups of users:

1. Users who work in phases of ED and EL and on one document at-a-time,
2. Users who work in phases of ED and EL and on a single hypothesis at-a-time, and
3. Users who do not work in phases of ED and EL.

When we categorised the users according to these non-exclusive groups, we found that six users fall into the first group, whereas five users fall into the second. Three users fall into both groups, namely User15, User16, and User20. Most users, seven to be exact, fall into the third group. This group also contains User2 and User3, who did not create any links between a piece of evidence and hypothesis, which means we cannot evaluate whether or not these users would have worked in phases of ED and EL. However, both users worked on one document at-a-time. Table 4.5 shows each group and the users who fall into it.

Group 1	Group 2	Group 3
User0, User5, User7, User15, User16, User20	User12, User14, User15, User16, User20	User1, User2, User3, User4, User17, User18, User19

Table 4.5: Distribution of the users in our study according to their behaviour.

To assess the correlation between the different kinds of behaviour, we calculated two metrics. First, Matthew’s correlation coefficient (Matthews, 1975) C , and second, the agreement A on all three possible combinations of variables. Baldi et al. (2000) defined Matthew’s correlation coefficient as

$$C = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}}, \quad (4.1)$$

where tp is the number of true positives, fp the number of false positives, and tn the number of true negatives. Its value range is $[-1, 1]$, with 1 indicating a perfect correlation, 0 indicating no correlation, and -1 indicating a perfect inverse correlation. We calculated the agreement between the different types of behaviours presented in table 4.4. Because the values are binary, we define the agreement as the accuracy of the labels in the columns.

We calculated the correlation and agreement between (1) phases of ED or EL and working on a single hypothesis at-a-time, (2) phases and working on a single document at-a-time, and (3) working on a single hypothesis and working on a single document at-a-time. We found that:

- (1) The correlation C of users who worked in phases and users who worked on a single hypothesis at-a-time is 0.595, with an agreement A of 0.75;
- (2) The correlation between phases and working on single documents at-a-time is -0.218 with an agreement of 0.438; and
- (3) Users who worked on single hypotheses showed a correlation of -0.234 , with an agreement of 0.313.

There are two reasons for the high correlation and agreement between users who worked in phases and on a single hypothesis at-a-time. First, all users who worked on a single hypothesis at-a-time also worked in phases; and second, 7 out of 16 users did not work

in phases or on a single hypothesis at-a-time. Also, all users who worked in phases but not on a single hypothesis at-a-time worked on a single document at-a-time. However, the opposite is not true, as there are users who worked on a single document at-a-time, but not in phases, e.g. User1 and User19. Generally, the users worked either on a single hypothesis at-a-time, or on a single document at-a-time, with User15 and User16 working both on a single hypothesis and single document. Only User4 did not work in phases, not on a single hypothesis at-a-time and not on a single document at-a-time.

We conclude from that that users who worked on a single hypothesis at-a-time are more likely to also work in phases, whereas users who work on one document at-a-time may or may not work in phases. Regarding users who worked neither on a single hypothesis, nor on a single document at-a-time, we find that this only applies to users User4 and User13, with User4 not working in phases of ED and EL, whereas User13 did.

4.3.5 Revisions of hypotheses

After the study, we asked the users whether they modified their hypotheses, finding that only a few users reported that they did. However, table 4.6 shows that all users, except User3, revised their hypotheses at least once. On average, the users modified their hypotheses 5 times, with User19 modifying their hypotheses 15 times and users User7, User13, and User15 modifying one hypothesis once.

User	Revisions
User0	9
User1	3
User2	10
User3	0
User4	5
User5	4
User7	1
User12	5
User13	1
User14	1
User15	3
User16	7
User17	5
User18	5
User19	15
User20	1

Table 4.6: Number of revisions of the hypotheses for each user.

A more detailed analysis showed that while most of these modifications were minor revisions and spelling corrections, some users modified their hypotheses to a greater extent. For instance, User16 started with the hypothesis “*Alternatives to nuclear energy are available*” and modified it to “*Alternatives to nuclear energy must be used*”. User19’s initial hypothesis was the statement “*No reversibility*” and through intermediate steps, they revised it to “*The risk of reversibility doesn’t get any attention*” and finally to “*The risk of reversibility probably doesn’t get any attention after events, if other factors are dominant*”. The last change shows a major revision because it makes the hypothesis much more specific compared with the previous version.

4.4 User Feedback

In the questionnaire, the users had to answer five multiple-choice and three free-form questions. Table 4.7 shows all multiple-choice responses from the questionnaire. The users were almost equally split regarding whether or not EDoHa was beneficial in substantiating their hypotheses, with 10 out of 16 reporting that it was. Only slightly more than half of the users reported gaining new insights (9 out of 16), and 11 users reported being willing to continue to use it in their studies. Although only eight users reported to have revised their hypotheses, the analysis of their behaviour showed that almost all users modified their hypotheses at least once. This is particularly apparent in User4, who reported to not have revised their hypotheses at all. However, the actual modifications of this user were not just minor revisions. For instance, the user first gave the hypothesis “*There is a call for alternative energy*” and then revised it to “*Renewable energy is the only alternative*”.

User	Substantiated	New Insights	Continue to use	Revisions	Work mode
User0	no	yes	yes	yes	Evidence first
User1	yes	yes	yes	yes	Evidence first
User2	yes	no	yes	no	Evidence first
User3	no	no	no	no	Hypothesis first
User4	no	no	no	yes	Evidence first
User5	no	yes	no	yes	Evidence first
User7	no	no	yes	no	Mixed
User12	yes	yes	yes	no	Evidence first
User13	yes	yes	yes	yes	Hypothesis first
User14	no	yes	no	yes	Evidence first
User15	yes	no	yes	no	Evidence first
User16	yes	no	no	yes	Evidence first
User17	yes	no	yes	no	Hypothesis first
User18	yes	yes	yes	no	Evidence first
User19	yes	yes	yes	yes	Hypothesis first
User20	yes	yes	yes	no	Evidence first
$10/16 = 62.5\%$ $9/16 = 56.25\%$ $11/16 = 68.75\%$ $8/16 = 50\%$					

Table 4.7: Responses of the users to the questionnaire.

Another disagreement between the behaviour the users reported and that we observed is in the users’ general approach. Four users reported having defined hypotheses first and then searching for evidence, but we only observed three users to have defined hypotheses in the beginning, namely User20, User13, and User19. Of those, only User13 and User19 also reported having started with a hypothesis, whereas User20 reported to have started with finding evidence, despite having specified multiple hypotheses in the beginning.

In the free-form feedback, we asked the users three questions: *what they liked*, *what they disliked*, and *what would like to have changed in EDoHa*. Regarding what the users liked, most of them responded with positive feedback regarding the clear layout and structure of EDoHa. Others mentioned that it enabled them to quickly gain an overview over the discourse. The most common negative point was the difficulty of selecting individual and complete sentences in the annotation editor. The users’ struggle is also visible in the number of annotations the users created in the beginning. Several users showed an increase of labelled pieces of evidence, with a subsequent decrease at the beginning of the study, indicating a struggle with the annotation editor. Because of the large number of users who mentioned this problem, we created an easier to use annotation editor for the second study. In the easier annotation editor, the user does not have to mark sequences

but instead can label a sentence by simply clicking on it. This reduced the learning curve involved in getting to use EDoHa.

4.5 Data

In our user studies, we created two datasets. To ascertain the possibility of creating a gold-standard corpus that would allow single ED and EL models to be useful for all users, we calculated the inter-annotator agreement of similar hypotheses from the nuclear energy dataset.

4.5.1 Corpus Description

We collected all the annotations from the users in our two studies to create our dataset, that is, all pieces of evidence each user labelled and the hypotheses they validated with them. The resulting datasets are described in table 4.8, where the number of sentences is the sum of the number of sentences of each document the user opened, how many pieces of evidence the user labelled, and where the two remaining columns show how many hypotheses the user formulated and how many links between hypotheses and evidence they created. If a hypothesis does not have at least one piece of evidence linked to it, it is not counted.

Topic	User	Sentences	Evidence	Hypotheses	Links
Nuclear Energy	User0	363	205	13	259
	User1	321	21	4	12
	User2	403	79	3	0
	User3	479	85	6	0
	User4	479	27	6	27
	User5	479	78	8	63
	User7	479	74	7	70
	User12	479	29	2	29
	User13	403	38	6	30
	User14	479	41	4	23
	User15	479	38	9	32
	User16	441	41	8	28
	User17	321	77	16	61
	User18	291	45	12	44
	User19	479	44	11	56
	User20	328	38	12	21
Forest Dieback	User0	700	124	8	183
	User1	508	34	6	45
	User2	700	72	3	69
	User3	700	47	9	61
	User4	700	40	4	19
	User5	603	52	4	26
	User6	700	26	6	26
	User7	433	46	9	46
	User8	700	154	0	0
	User9	603	37	5	12
	User10	700	25	4	29
	User11	569	32	5	32
	User12	700	56	0	0

Table 4.8: The number of evidence annotations, hypotheses, and links each user created.

In the nuclear energy dataset, the data for each user vary greatly. Some users defined many hypotheses, e.g. 16 in the case of User17 and 13 in the case of User0, and others, such as User12, defined only two. On average, the users formulated ≈ 8 hypotheses and linked ≈ 41 pieces of evidence to them. In both dataset created User0 considerably more data, i.e. in labelled sentences and links between hypotheses and evidence, than any other user. Two users, User2 and User3, linked no evidence to any hypothesis. We think that this is because of a misunderstanding of the purpose of the study in the case of User2, who believed it to be a usability study. In the case of User3, we think they might have missed the Drag & Drop functionality to create links between hypotheses and evidence in the introduction of EDoHa.

In the forest dieback dataset, the users generally formulated fewer hypotheses (≈ 4) but linked a similar number of pieces of evidence to them (≈ 42). This dataset also contains two users who created no link between hypotheses and pieces of evidence, namely User8 and User12.

4.5.2 Inter-Annotator Agreement

We calculated the inter-annotator agreement of the nuclear energy dataset by first creating pairs of similar hypotheses and then calculating the agreement between the sentences labelled as evidence. This shows how well users agree on evidence; if the agreement is high, then we can create a gold-standard dataset for ED for researchers in the humanities and social sciences, and a single model would be able to learn to find evidence for all users. If the agreement is low, then we cannot create a gold-standard dataset and each user would need to have an individualised ED model. Furthermore, we calculated the agreement of all pairs of hypotheses from different users to see whether different users used the same evidence to support different hypotheses.

To calculate the agreement for similar hypotheses, we first asked an expert annotator in history to select pairs of hypotheses for which they would expect an overlap in evidence, that is overlap_e . We then calculated the Cohen's κ (Cohen, 1960) of the annotations at the sentence level. We limited the documents on which we calculated the agreement to the ones from which either user extracted the evidence. We calculated the agreement of the two hypotheses h_u and h_v by creating two copies of all source documents d_u and d_v , and in each of them, we labelled only the pieces of evidence also linked to the corresponding hypothesis. This means that in d_u , we labelled only those pieces of evidence linked to h_u , and in d_v , we labelled the pieces of evidence linked to h_v . If a user did not open a particular document, we treated it as if it contained no pieces of evidence.

Because of the large differences in class distribution, which is visible in the difference between the number of sentences and pieces of evidence for each user in table 4.8, hence causing issues with Cohen's κ (Artstein and Poesio, 2008), we decided to also look at the overlap in evidence. Formally, we define the overlap of the evidence of two hypotheses $\text{ev}(h_u)$ and $\text{ev}(h_v)$ as

$$\text{overlap}_e(h_u, h_v) = \frac{|\text{ev}(h_u) \cap \text{ev}(h_v)|}{|\text{ev}(h_u) \cup \text{ev}(h_v)|} \quad (4.2)$$

We also created pairs of hypotheses from different users and calculated the agreement of their evidence to see whether users support different hypotheses with the same evidence. Finally, in a separate study, we had two users search for evidence related to three pre-defined hypotheses. The separate study was conducted on a different group of students but with the same topic of nuclear energy and political speeches as textual sources. Furthermore, we selected the hypotheses from the ones the users created in our first study to

make sure that the participants could find evidence to evaluate them. Table 4.9 shows the agreement of the pairs of hypotheses the expert annotator deemed similar and the pairs of hypotheses that showed a substantial agreement with their evidence, i.e. $\kappa > 0.6$ (Landis and Koch, 1977).

Hypotheses pair		κ	e
International security arrangements in the nuclear sector are necessary	Nuclear power and security: further expansion of domestic and foreign policy	0.116	0.067
Nuclear-phase-out is not possible due to the profit motive of corporations	Profit maximisation of the economy	0.067	0.042
Chernobyl as a reminder for the nuclear-phase-out	Chernobyl and Fukushima repeatedly related	0.057	0.042
Nuclear phase-out should not be slowed down by individual companies	Is money and the economy put on the safety of each one?	-0.007	0.000
Does the nuclear industry have too much power?	Criticism of Fukushima	1.000	1.000
Security of nuclear reactors must be guaranteed	The following security measurements	0.748	0.600
Does the nuclear industry have too much power?	If the information policy comes from one actor, there is a high probability that not all information will reach the public	0.666	0.500
Criticism of Fukushima	If the information policy comes from one actor, there is a high probability that not all information will reach the public	0.666	0.500
The electrical grid has to be restructured to accommodate the nuclear phase-out		-0.030	0.000
The nuclear phase-out is fundamentally connected to the industry		0.033	0.033
The risks of nuclear energy do not stop at national borders		0.037	0.034

Table 4.9: Agreement of the evidence of similar hypotheses (top), all hypothesis pairs with a substantial agreement on evidence (middle), and agreement of pre-defined hypotheses (bottom)

We found that similar hypotheses show little to no agreement. For instance, the hypotheses “*International security arrangements in the nuclear sector are necessary*” and “*Nuclear power and security: further expansion of domestic and foreign policy*” only show an agreement of 0.116κ and 0.067overlap_e . The hypotheses “*Chernobyl as a reminder for the nuclear-phase-out*” and “*Chernobyl and Fukushima repeatedly related*” also show a low level of agreement with 0.057κ and 0.042overlap_e .

Among hypotheses with a substantial agreement, we found that many of them are not similar. Some are related, as in having a similar aspect, e.g. *security*, but others are completely dissimilar. For instance, the hypotheses “*Security of nuclear reactors must be guaranteed*” and “*The following security measurements*” show a substantial agreement

with 0.748κ and 0.6overlap_e , which might be because both are related to *nuclear safety*. The two hypotheses with a perfect agreement are “*Does the nuclear industry have too much power*” and “*Criticism of Fukushima*”. This also shows that these particular users interpreted the same statements differently. Although the first one interprets them as pointing out that there are power dynamics, the second one interprets them as general criticism.

When providing the users with hypotheses, e.g. “*The electrical grid has to be restructured to accommodate the nuclear phase-out*”, we found no agreement with -0.030κ and 0.000_e . The two other hypotheses, namely *The nuclear phase-out is fundamentally connected to the industry* and *The risks of nuclear energy do not stop at national borders* also show no agreement, with $\approx 0.35\kappa$ and $\approx 0.33\text{overlap}_e$. This means that the disagreement is not based on having dissimilar hypotheses because we found no agreement, even when we provided the users with hypotheses. We conclude from this that the creation of a gold-standard dataset that fits all users perfectly is not possible.

4.6 Discussion

We are aware of a few points of criticism of the findings in this chapter, most importantly, regarding the agreement of evidence on similar hypotheses and the potential influence of the user interface of EDoHa on the individual approach for finding evidence and validating hypotheses.

Agreement of evidence A general approach in NLP to create large amounts of annotated data is to use crowdworkers. The crowdworkers are given a description of their task and a set of criteria that have to be fulfilled, e.g. that an argument expresses a reasoning step or is a piece of evidence (Stab et al., 2018b). The same data are then presented to multiple crowdworkers and later aggregated with different methods, e.g. majority voting. However, the annotation of such data is oftentimes intentionally context-independent and, therefore, not realistic. Although every person might agree that a particular sentence is a piece of evidence related to a particular hypothesis, not every person would select it from a collection of possible candidates. The particular sentence a user selects might depend on the order in which they read a document, their previous exposure to other candidate sentences, or simply the personal preference that one sentence expresses a particular point more clearly. Considering that our goal is supporting the individual user, we have to take these personal preferences into account, so we cannot use any aggregation method because it would reduce the amount of support any method can give to the individual user.

Phases of ED and EL Although the user interface of EDoHa facilitates working in phases by offering two distinct views—one for a document and one for the hypotheses—we found a higher correlation of users who worked in phases and those who worked on a single hypothesis at-a-time than users who worked on a single document at-a-time. Given the different view for documents, we would have expected that the users first read a document and then after they had finished, they would switch to the evidence linking view and link all found pieces of evidence to hypotheses; in this way, the users would not need to look into multiple documents and only switch once they finished reading one. However, several users who worked on one document at-a-time did not work in phases. They would link each piece of evidence to a hypothesis once they found it and then continue reading the same document. A possible reason is the high cognitive load required to keep multiple different hypotheses in mind while searching for evidence. This can possibly be addressed

by enabling the user to choose a hypothesis to link a newly labelled piece of evidence to, or by improving the order in which evidence appears in the list of evidence at the bottom of the evidence linking view. Another possible reason for this is that particular approaches have been learned by users. For instance, if a user already gained experience in existing annotation tools that enable directly assigning a code to a piece of text, they might continue to prefer using the previously learned approach.

4.7 Chapter Conclusion

In this chapter, we addressed four sub-questions whose answer will lay out the basic requirements for supporting researchers in the humanities and social and political sciences to find evidence and validate their hypotheses.

Regarding our sub-questions, we found the following:

- ① *Do researchers distinguish between finding evidence and linking it to hypotheses?*

There is no canonical user after which ED and EL approaches can be modelled. In a user study with 16 users, we saw different approaches that varied from users who first collected all pieces of evidence and then linked them to hypotheses to users who once they found a piece of evidence immediately linked it to a hypothesis. However, it is possible that the user interface of EDoHa influenced the users. The strong distinction between ED and EL with separate views might have facilitated a more phased approach.

- ② *Do they read a document multiple times or just once?*

We found that the majority of users worked on one document at-a-time and did not return to it after they finished reading one document. We also found that the change from the ED phase to the EL phase did not generally occur after the user finished reading a document.

- ③ *Do they work on multiple hypotheses in parallel or one at-a-time?*

Although some users held multiple hypotheses in their mind at the same time, others worked on only one hypothesis at-a-time. Once they validated a hypothesis sufficiently, they did not return to it.

- ④ *Do researchers revise their hypotheses over time, thus developing them, or are the hypotheses defined once in the beginning and then remain static?*

Many users modified their hypotheses throughout the study. Although many of these modifications were corrections in orthography, other modifications show that the users developed their hypotheses over time and updated them as they found more evidence.

- ⑤ *How well do they agree on the evidence?*

We found that different users used different pieces of evidence to evaluate similar hypotheses and support vastly different hypotheses with similar to identical evidence.

From the first two findings, we can conclude that an annotation tool may not force the user to use one particular approach, but it must be flexible enough to enable the user to follow their own personal approach. From the other findings, any ED and EL model aiming at supporting researchers must not only be specific to an individual user, but because of the changing nature of the user's goals, it must adapt dynamically to the current user's

needs. We also found three groups of users by which we can categorise the users: first, those who work in phases of ED and EL and on one document at-a-time; second, those who work in phases of ED and EL and on one hypothesis at-a-time; and the users who did not work in phases but often on one document at-a-time. This investigation also provides information regarding possible simulation strategies. For instance, the first group worked in phases and on one document at-a-time, which we can simulate for ED experiments. Similarly, for EL, our finding that most users work on multiple hypotheses in parallel means a cross-validation setup is a more realistic simulation than a cross-hypothesis setup.

A future direction of research can be on the influence of an individual user's background on the particular work approach. Because we have users from different backgrounds, such as history, political science, and social science, this background might also bias the formulation of hypotheses and selection of evidence. Although this might be an additional factor, we could not assess it because the study was conducted in an anonymous fashion, so we did not collect data regarding an individual user's background. Another possible source of bias is the level of experience of the individual user. For instance, an early bachelor student might select different sentences than a post-doc. However, such an analysis is not possible within the constraints of a teaching-seminar while retaining anonymity.

Chapter 5

Machine Learning for Evidence Detection

There are many different ways in which ED methods can be used to support researchers in finding evidence. Although we have established that any ED method must adapt to its user, it is still unclear which approach is the best. However, we first need to define ED for the purpose of this chapter.

Definition 1. Let s be a sentence in a document \mathcal{D} and $\text{ev}(s)$ the predicate of a sentence being a piece of evidence. Then, Evidence Detection is a binary classification function $f \rightarrow \{0, 1\}$ so that we can find all pieces of evidence within the document $\{\forall s \in \mathcal{D} : \text{if ev}(s)\}$. The goal of this task is to learn this binary sentence-classification function f .

There are three basic approaches we can follow. First, we could train a model directly on the data a user generates. Second, the differences between the users might be within the generalisation abilities of a well generalising, state-of-the-art method trained on user-independent data. Third, we could use the out-of-domain data to pre-train an ED model and then fine-tune it towards a specific user. Furthermore, if an interactively trained or fine-tuned model performs better than a well generalising one trained on out-of-domain data, it might require a minimum amount of interactive training data to do so. We therefore also compare the user-specific models against a *direct transfer* approach with a well generalising, state-of-the-art model. This leads us to formulate the following sub-questions to answer our second research question (RQ 2: How well do machine learning-based methods work for ED?), which figure 5.1 illustrates:

- ① Does a user-specific ED model trained on small, in-domain data outperform a well generalising, user-independent one trained on out-of-domain data?
- ② Can we use user-independent, out-of-domain data for pre-training to reduce the necessary amount of user-specific, in-domain training data?
- ③ How much training data does a user have to generate so that the interactively trained models can outperform a well generalising, user-independent model?
- ④ How close to the user-specific data must the training data of the user-independent model be, until it performs equally well as the user-specific one?

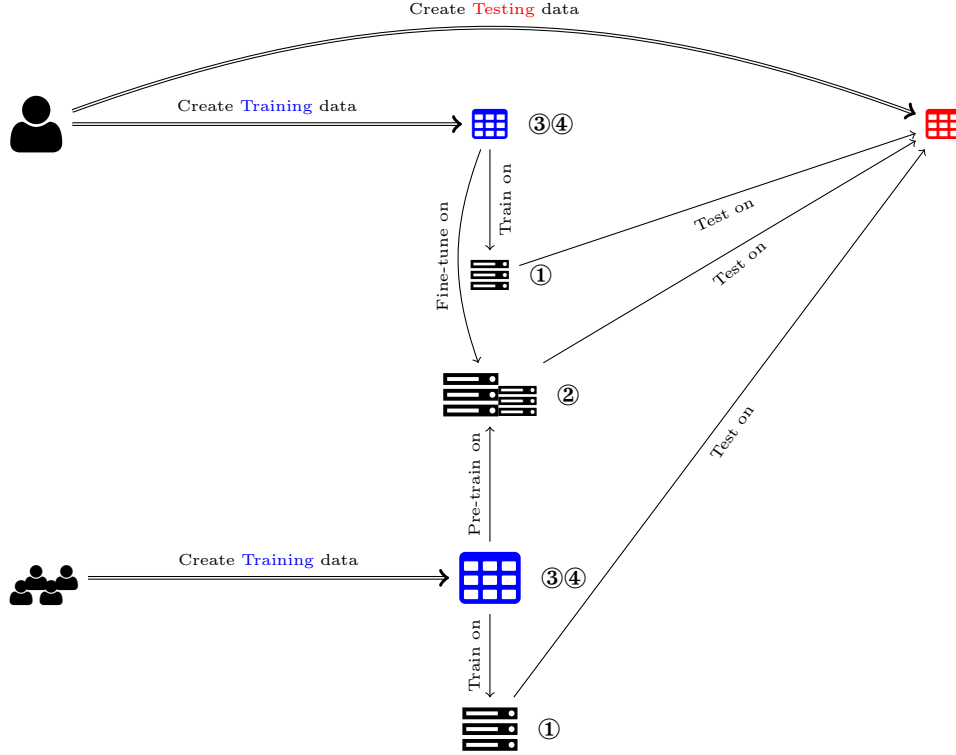


Figure 5.1: Overview of how the user-dependent, in-domain data and user-independent, out-of-domain data relate to models and the different sub-questions.

We address these questions in the following steps. First, we evaluated how well an ED model, which was trained directly on the user-generated data, performs. Second, we compared different kinds of transfer learning, such as *direct transfer* and *fine-tuning*, with the previous *direct training* approach. Third, we created a more realistic user simulation than random sub-sampling which is currently used to determine the necessary amount of training data (e.g. Schulz et al., 2018; Schulz et al., 2019). Fourth, we compared the two interactively trained models, one directly trained, and one fine-tuned, with a well generalising, state-of-the-art model that was trained on user-independent or out-of-domain data. To address the fourth question, we employed an additional dataset and used a cross-topic evaluation to evaluate the different transfer learning approaches. In this evaluation, the domain shift is much smaller than in the previous ones.

5.1 Direct Training

To investigate whether an ED model can be trained on the nuclear energy and forest dieback datasets described in chapter 4, we evaluated the performance of several machine learning techniques and compared them against a majority and random baseline.

5.1.1 Models

We investigated six different models: four models learn to classify sentences as evidential or not purely based on sample sentences created by the user; two models also consider the hypotheses the users created. We then compared these models against the baselines to evaluate their quality.

Baselines

Majority The majority baseline always predicts the majority class for each user. If a user labelled less than half the sentences in the training data as evidence, then this model will never predict any candidate as evidence; if a user labelled more than half the sentences as evidence, then it will predict all candidates as evidence.

Random The random baseline first determines the class distribution of the training data. For predictions, it samples randomly from this distribution.

Hypothesis-Agnostic Models

Of the four hypothesis-agnostic models we evaluated, two use word n-grams as input, namely the Naïve Bayes and the Multi-Layer Perceptron (MLP). The other two models use 100-dimensional word embeddings published by Reimers et al. (2014) as input.

Naive Bayes The Naive Bayes classifier uses bag-of-words as input features. The pre-processing first removes the stopwords based on a list provided by NLTK¹ and then determines the word bi-grams. The implementation is based on scikit-learn².

MLP The MLP uses the same features as the Naive Bayes classifier, except with word uni-grams rather than bi-grams. Similarly, the implementation is based on scikit-learn.

bilstm(50) This model uses a Bi-directional Long-Short Term Memory (bilstm) with 50 hidden nodes and a dense layer for classification. To account for the class imbalance, it weighs the classes similar to King and Zeng (2001) through the implementation in scikit-learn. The input features are the word embeddings published by Reimers et al. (2014).

bilstm(100) We implemented this model to investigate, whether the reduction in dimensionality from 100-dimensional word embeddings to 50 nodes in the lstm layer caused a loss in performance. It has as many nodes in the lstm layer, as the input has features, so it does not cause any potential loss due to the feature compression.

Hypothesis-Aware Models

The hypothesis-aware models predict whether a sentence fits to one of the user-defined hypotheses $h \in \mathcal{H}$. Both models use the average word embeddings of a candidate sentence and one or more hypotheses to classify the candidate sentence as evidence or not. The word embeddings are identical to the ones used for the bilstms.

cos(s, h) The semantic similarity based model calculates the similarity between a candidate sentence s and a hypothesis h by calculating the cosine-similarity of their average word embeddings. This is then repeated for each hypothesis the user created. This means a sentence s is predicted as evidence if $\exists h \in \mathcal{H} : \cos(s, h) > \tau$. Before averaging the sequence of word embeddings, the stopwords are removed based on the list provided by NLTK. Much like the bilstms, this model is based on Keras.

¹<https://www.nltk.org/>

²<https://scikit-learn.org/stable/>

link(s, h) This classifier aims at finding evidence by predicting to which hypothesis a candidate sentence would be linked. This means that it uses the links the user created between evidence and hypotheses as training data and then predicts whether or not a candidate sentence would be linked to a particular hypothesis. Formally defined, it predicts a sentence s to be evidence if $\exists h \in \mathcal{H} : \text{link}(s, h)$. In training, the linking model first loads all links between the hypotheses and pieces of evidence from the training documents. It then creates an equal amount of non-links by randomly pairing hypotheses with pieces of evidence from the training data. It then filters out all user-created links that are present in the random pairs. To allow for filtering, it creates 10 times as many random pairs as there are links in the data; the number of non-links is then down-sampled to the number of links to have a balanced dataset. The model is then trained to predict whether a candidate sentence / hypothesis pair is linked or not. Both vectors are concatenated before being fed into the network. For prediction, a candidate sentence is paired with all hypotheses. If the model predicts that the candidate sentence would be linked to at least one hypothesis, it returns that the candidate sentence is a piece of evidence. This model is also implemented in Keras³.

Hyper-parameter Optimisation

We conducted hyper-parameter tuning on a development user, namely User7 in the nuclear energy dataset. We conducted no separate hyper-parameter tuning for the forest dieback dataset. We selected this user because they created slightly above average annotations and links. Table 5.1 shows the hyper-parameters for the MLP, bilstm(50), bilstm(100), and link(s, h).

Model	HL	HLS	LR	Ep	BS	D
MLP	1	10	0.001	50	32	1.0
bilstm(50)	1	50	0.001	10	32	0.5
bilstm(100)	1	100	0.001	10	32	0.5
link(s, h)	3	(100, 75, 50)	0.01	250	32	0.5

Table 5.1: Hyper-parameters used for the direct training evaluation. The hyper-parameters are the number of hidden layers (HL), the size of the hidden layers (HLS), the learning rate (LR), the number of epochs (Ep), the batch size (BS), and the dropout (D).

We used the same user to determine the threshold τ of the $\cos(s, h)$ model. We evaluated different values and found the best performance with the threshold $\tau = 0.7$.

5.1.2 Evaluation

We based our evaluation on the finding that most users worked on one document at-a-time and once they finished reading it, they did not return to it. This means, we used a document based evaluation setup and conducted the experiments for each user separately. More specifically, we selected a leave-one-document-out setup because of a small number of separate documents. For models that contain stochastic components, i.e. neural networks and the random baseline, we conducted each experiment with five different randomisation seeds, namely $s \in \{0, 1, \dots, 4\}$. For the nuclear energy dataset, we used all models, and for the forest dieback dataset, we used the three best-performing models in the evaluation.

Table 5.2 shows the results averaged across the individual users. The bilstm(100) reached the highest macro-averaged F1 score for all models and the highest macro-averaged

³<https://keras.io/>

recall. For both the macro-averaged and evidence class precision, the MLP reached the highest scores. The $\cos(s, h)$ model reached by far the highest recall on the evidence class and, therefore, also the highest evidence-F1 score.

		F1	P	R
Evidence & No Evidence	Majority	0.462 (0.032)	0.433 (0.049)	0.500 (0.000)
	Random	0.491 (0.013)	0.491 (0.012)	0.491 (0.013)
	NaiveBayes	0.501 (0.005)	0.506 (0.005)	0.507 (0.007)
	MLP	0.513 (0.011)	0.538 (0.001)	0.516 (0.007)
	bilstm(50)	0.527 (0.000)	0.527 (0.001)	0.533 (0.001)
	bilstm(100)	0.528 (0.000)	0.528 (0.001)	0.534 (0.002)
	$\cos(s, h)$	0.271 (0.166)	0.505 (0.040)	0.508 (0.182)
	$\text{link}(s, h)$	0.387 (0.048)	0.445 (0.102)	0.444 (0.051)
Evidence only	Majority	0.051 (0.186)	0.040 (0.145)	0.071 (0.258)
	Random	0.126 (0.132)	0.127 (0.131)	0.126 (0.133)
	NaiveBayes	0.169 (0.014)	0.151 (0.059)	0.202 (0.031)
	MLP	0.132 (0.000)	0.215 (0.030)	0.104 (0.028)
	bilstm(50)	0.202 (0.004)	0.183 (0.015)	0.231 (0.003)
	bilstm(100)	0.202 (0.001)	0.185 (0.006)	0.229 (0.011)
	$\cos(s, h)$	0.217 (0.410)	0.152 (0.121)	0.768 (0.424)
	$\text{link}(s, h)$	0.123 (0.028)	0.091 (0.040)	0.300 (0.138)

Table 5.2: Results of the nuclear energy dataset. The values for the evidence and no evidence rows are macro-averaged.

When broken down to an individual user, the results vary greatly, as table 5.3 illustrates.⁴ Although the $\cos(s, h)$ model reached the highest evidence-F1 score more often than the other models, the results are heavily dependent on the individual user. For instance, all models except the linking model reached their highest evidence-F1 score for User0. For User3, neither the linking nor the semantic similarity models reached any result, which is understandable, because this user did not create any link between the evidence and hypotheses.

We limited the evaluation of the forest dieback dataset to the three models that performed best across all users: bilstm(50), bilstm(100), and $\cos(s, h)$. Table 5.4 shows that the performance measured as the macro-averaged score across both classes increased compared with the nuclear energy dataset. However, the evidence-F1 score for both bilstm(50) and bilstm(100) dropped by about 0.04; the same drop can be found on the evidence-precision. The recall for both models dropped by 0.03 (bilstm(50)) and 0.013 (bilstm(100)), respectively. The $\cos(s, h)$ model, on the other hand, increased its recall by around 0.008 but dropped considerably in precision by 0.067, cutting it almost in half.

For each individual user (table 5.5), we find that the performance in evidence-F1 score is generally lower than for the nuclear energy data. However, they generally outperform the random baseline.

5.1.3 Discussion

In our evaluation we identified three particularly surprising observations: the high performance of the $\cos(s, h)$ model in the nuclear energy dataset, the performance drop between

⁴Our results differ from the ones presented by Stahlhut, Stab, and Gurevych (2018) due to a different approach in calculating the macro-F1 score. Stahlhut, Stab, and Gurevych (2018) calculated the macro-F1 score from the macro-averaged precision and recall, whereas we calculate it from the class-specific F1 scores.

Evidence F1 score							
	Random	NaiveBayes	MLP	link(s, h)	cos(s, h)	bilstm(50)	bilstm(100)
User0	0.559	0.511	0.549	0.009	0.731	0.641	0.630
User1	0.028	0.143	0.054	0.077	0.134	0.025	0.069
User3	0.173	0.169	0.151	0.000	0.000	0.269	0.228
User4	0.014	0.071	0.144	0.054	0.107	0.104	0.095
User5	0.139	0.212	0.170	0.235	0.310	0.245	0.263
User12	0.033	0.047	0.000	0.022	0.127	0.057	0.060
User13	0.091	0.144	0.031	0.160	0.109	0.184	0.155
User14	0.103	0.092	0.039	0.121	0.208	0.149	0.150
User15	0.083	0.229	0.180	0.059	0.155	0.244	0.226
User16	0.128	0.103	0.109	0.146	0.181	0.118	0.128
User17	0.226	0.333	0.280	0.278	0.371	0.354	0.363
User18	0.112	0.190	0.040	0.347	0.288	0.104	0.102
User19	0.096	0.099	0.020	0.053	0.171	0.174	0.175
User20	0.058	0.021	0.078	0.154	0.152	0.165	0.187

Table 5.3: The values for each model are the evidence-F1 score for the nuclear energy dataset.

		F1	P	R
Evidence & No Evidence	Majority	0.479 (0.007)	0.461 (0.036)	0.500 (0.014)
	Random	0.494 (0.001)	0.494 (0.000)	0.494 (0.000)
	bilstm(50)	0.540 (0.012)	0.537 (0.007)	0.551 (0.001)
	bilstm(100)	0.534 (0.018)	0.532 (0.010)	0.548 (0.001)
	cos(s, h)	0.247 (0.195)	0.530 (0.144)	0.553 (0.068)
Evidence only	Majority	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	Random	0.069 (0.000)	0.068 (0.001)	0.071 (0.001)
	bilstm(50)	0.166 (0.031)	0.142 (0.039)	0.204 (0.004)
	bilstm(100)	0.162 (0.054)	0.132 (0.051)	0.216 (0.013)
	cos(s, h)	0.150 (0.445)	0.085 (0.349)	0.844 (0.224)

Table 5.4: Results of the forest dieback dataset. The values for the evidence and no evidence rows are macro-averaged.

Evidence F1 score				
	Random	cos(s, h)	bilstm(50)	bilstm(100)
User0	0.158	0.326	0.320	0.347
User1	0.047	0.222	0.052	0.074
User2	0.079	0.202	0.222	0.246
User4	0.041	0.118	0.183	0.174
User5	0.090	0.173	0.175	0.165
User6	0.054	0.074	0.058	0.025
User7	0.107	0.223	0.300	0.270
User9	0.064	0.124	0.183	0.170
User10	0.023	0.075	0.186	0.148
User11	0.047	0.112	0.047	0.053
User12	0.045	0.000	0.104	0.110

Table 5.5: The values for each model are the evidence-F1 score for the forest dieback dataset.

the nuclear energy and forest dieback datasets, and the lack of comparability between the two datasets.

Reasons for the high performance of $\cos(s, h)$ When analysing the performance of the $\cos(s, h)$, we found all users for which it worked particularly well; User0, User5, and User18, all mentioned Fukushima or Chernobyl in more than one hypothesis. After an investigation into the user-defined hypotheses, we hypothesised that the users for which the $\cos(s, h)$ model performed comparatively well used terms that also often occur in the pieces of evidence. We then calculated the lexical overlap between the existing links of the hypotheses and pieces of evidence the users created, comparing this with the lexical overlap of users with a particularly low performance of the model, namely User3, User4, and User13. However, we found no differences in lexical overlap between the hypotheses and pieces of evidence for users for which the model performed well and those for which it did not. We also hypothesised that the users for whom the model worked well had a higher number of hypotheses, giving the model more options to predict a link. A deeper investigation led to the finding that if the number of hypotheses is four or less, the $\cos(s, h)$ does not perform very well. If the number of hypotheses is higher, on the other hand, the performance increases. On the nuclear energy dataset, this is apparent in the cases of User0, who formulated 13 hypotheses, and User13, who formulated 12; in both cases the model performed comparatively well. User0 formulated six hypotheses, but the model could not predict any piece of evidence correctly. Although this finding might explain the high performance of the model, it does not mean that there is no exception. User12 formulated only 2 hypotheses, but the model still performed well. Therefore, we conclude that a large number of hypotheses is beneficial, but not sufficient for a high performance of the $\cos(s, h)$ model.

Performance drop between the nuclear energy and forest dieback datasets We found that the performance still dropped considerably, particularly in the precision of the evidence class. This is most likely because of a bug in EDoHa during the second user study in which opening another document required changing the view. This caused multiple users to first click on several documents that did not open until we pointed out the bug and how to work around it. Therefore, many users opened multiple documents without reading them, causing the documents to be included in the dataset without any annotations. This also caused these documents to be included in the training data, meaning that there were potentially sentences that a user would have labelled as evidence creating inconsistent training data. To test this hypothesis, we removed all files which a user did not open for more than 20 seconds from the test data. Table 5.6 shows the results on the evidence class after applying the time filter. Although all models, except for the majority baseline, improved in evidence-F1 score and especially in evidence-precision, the difference is almost always below 0.01, indicating that the filter duration was too small to have an effect. However, the recall of all models increased. This means that the time filter removed documents in which the users labelled sentences as evidence which the models did not predict. We therefore conclude that although the users opened many documents in the beginning, they still read them and annotated pieces of evidence in them.

Lack of comparability of individual users across datasets Although desirable, it is not possible to evaluate how well a model performed for a particular person in both datasets. First, the participants were not assigned their seat in our computer lab, but selected it on their own and the account credentials were handed out based on the seating

		F1	P	R
Evidence only	Majority	0.000 [+0.000]	0.000 [+0.000]	0.000 [+0.000]
	Random	0.071 [+0.002]	0.072 [+0.003]	0.070 [+0.000]
	bilstm(50)	0.172 [+0.006]	0.151 [+0.009]	0.206 [+0.002]
	bilstm(100)	0.167 [+0.005]	0.139 [+0.007]	0.217 [+0.001]
	$\cos(s, h)$	0.160 [+0.010]	0.092 [+0.007]	0.849 [+0.005]

Table 5.6: Results of the forest dieback dataset after filtering the test documents that have been open for less than 20 seconds. The values in brackets are the difference with the results without the time filter.

arrangement. Second, being included in a teaching-seminar meant that not every participant was present in both studies. The students were allowed to miss a predefined number of seminar classes, and these also included participation in our study. Third, the study was conducted anonymously; disclosing any information that could identify a participant would have severe consequences. For instance, disclosing that a participant was present in only the second study might make this participant identifiable.

5.2 Transfer Learning

We investigated two different kinds of transfer learning: *direct transfer* and *fine-tuning*, and compared their applicability with the previously evaluated *direct training* approach. In the direct training approach, we trained a model directly on a small amount of in-domain data. In both the direct transfer and fine-tuning approaches, we used large amounts of external, out-of-domain data for training. In the case of the direct transfer approach, we applied the trained model without further modifications to the in-domain data. For fine-tuning, we first performed similar training on the out-of-domain data but conducted additional fine-tuning on the in-domain data. Figure 5.2 shows the different approaches and their domains.

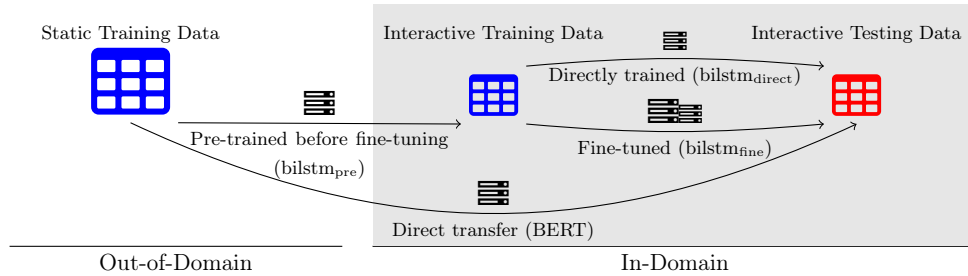


Figure 5.2: The relation between the out-of-domain and in-domain datasets and different training setups.

5.2.1 Datasets

We used three datasets in our evaluation: two for the task of ED and one AM dataset to investigate how well our results generalise to related tasks. We used the *ED-ACL-2014* (Aharoni et al., 2014) and *ED-EMNLP-2015* (Rinott et al., 2015) as the in-domain ED dataset. As the out-of-domain dataset, we used *ED-ACL-2018* (Shnarch et al., 2018). As the AM dataset, we used the *UKP Sentential AM* dataset published by Stab et al. (2018b).

In-Domain Datasets

ED-ACL-2014 This dataset consists of two parts. First, it contains a collection of claims from Wikipedia articles related to 33 different topics, and second, there are 12 topics with evidence related to 350 claims. The topics are randomly selected debate motions from the debate portal of the *International Debate Education Association*⁵. The dataset was constructed by five annotators who randomly selected a topic and then searched for suitable articles on Wikipedia. They then labelled candidates for claims relevant to the topic. These candidates were then cross-examined by other annotators. This process was then repeated for all 12 topics and 350 accepted claims to find evidence that can be classified as either of three classes: *study data*, which refers to quantitative results, *expert opinions*, and *anecdotes*. The annotated pieces of evidence are not limited to individual sentences, but could also span multiple sentences; in one case, a piece of evidence is 16 sentences long. This dataset contains no split into training, development, or testing data. We limited our experiments to the ED part of the dataset with 12 topics. Table 5.7 shows the topics and how many documents and pieces of evidence the dataset contains. It also shows that not only the absolute number of documents or pieces of evidence varies greatly between the topics, but also the average number of pieces of evidence per document. Figure 5.3 shows the variation in the length of the pieces of evidence in the dataset. About half the pieces of evidence are one sentence long, and the vast majority is less than four sentences in length. Because we define ED as sentence classification task, we need to convert the pieces of evidence that are more than one sentence long into individually labelled sentences.

Topic Id	Topic	Docs	Evidence
0	All collective bargaining rights claimed by trade unions	6	6
1	Atheism is the only way	19	124
2	Boxing	10	27
3	Gambling	11	74
4	Make physical education compulsory	13	35
5	Subsidise poor communities	10	46
6	That the right to asylum should not be absolute	13	71
7	The United States is responsible for Mexico's drug wars	6	33
8	The one child policy of the republic of China	13	77
9	The sale of violent video games to minors	15	132
10	The use of affirmative action	20	134
11	The use of performance enhancing drugs in professional sports	7	36

Table 5.7: Topic ids, topic titles, number of documents, and number of pieces of evidence in the ED-ACL-2014 dataset.

ED-EMNLP-2015 This dataset is an extension of the previous ED-ACL-2014 dataset. It contains 46 additional topics and additional annotations on the original 12 topics. Furthermore, the topic is stated as a *debate motion*, meaning it often starts with the prefix *This house would* or *This house believes that*. The dataset is split into 19 training and development topics and 39 training and testing topics. When conducting a leave-one-topic-out evaluation, the 19 training and development topics can be used for hyper-parameter tuning and the 39 training and testing topics for the final evaluation. Statistical information about this dataset can be found in table 5.8. Figure 5.4 shows the variation in the lengths in the pieces of evidence in the dataset.

⁵<http://idebate.org/>

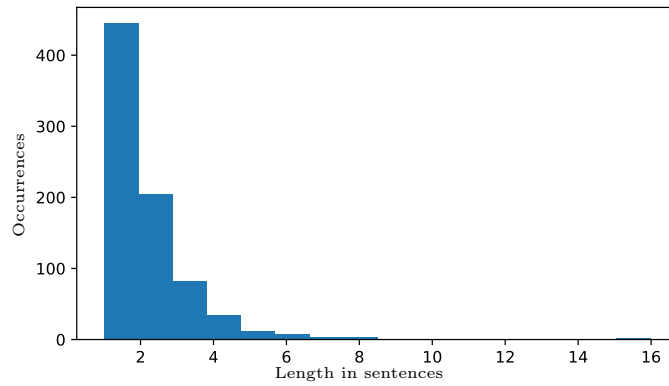


Figure 5.3: Histogram detailing the variation in the lengths of evidence in the ED-ACL-2014 dataset.

Split	Topics	Docs	Evidence
train and dev	19	170	857
train and test	39	377	1742

Table 5.8: Statistics on the ED-EMNLP-2015 dataset.

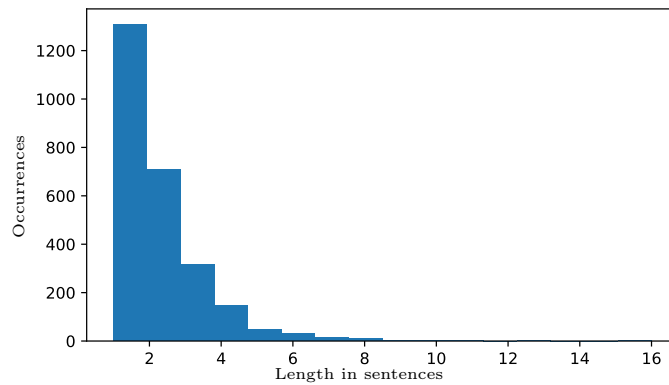


Figure 5.4: Histogram detailing the variation in the lengths of evidence in the ED-EMNLP-2015 dataset.

We removed the topics that overlap with the ED-ACL-2014 dataset from the ED-EMNLP-2015 dataset to keep the datasets separate. Table 5.9 shows the topic, number of documents, and number of pieces of evidence in the final test data of the ED-EMNLP-2015 dataset.

Out-Of-Domain Dataset

ED-ACL-2014 This dataset is distinct from the other ED datasets in that it does not contain any claims and that the pieces of evidence are limited to individual sentences. Furthermore, the data were not annotated by five expert annotators but by 10 crowdworkers who labelled each sentence. It covers 118 topics, of which 83 were selected as training data and 35 for testing. The pieces of evidence were also extracted from Wikipedia articles. A third difference is that this dataset does not provide the original content of the source articles from where the evidence was extracted. Therefore, we use this dataset as

Topic Id	Topic	Docs	Evidence
0	This house believes that Europe should weaken its austerity measures to guarantee its citizens greater social support	5	15
1	This house believes that Israel should lift the blockade of Gaza	11	168
2	This house believes that bribery is sometimes acceptable	4	6
3	This house believes that countries with an imbalanced male/female ratio skewed towards males should encourage parents to produce girls	4	5
4	This house believes that democratic governments should require voters to present photo identification at the polling station	3	4
5	This house believes that endangered species should be protected	13	18
6	This house believes that housewives should be paid for their work	2	3
7	This house believes that male infant circumcision is tantamount to child abuse	14	146
8	This house believes that opinion polls harm the democratic process	4	10
9	This house believes that the Church of England should be separated from the British state	4	17
10	This house believes that wind power should be a primary focus of future energy supply	17	35
11	This house believes the US is justified in using force to prevent states from acquiring nuclear weapons	8	30
12	This house prefers trade to aid	15	55
13	This house would abolish intellectual property rights	9	29
14	This house would abolish the monarchy	12	39
15	This house would ban all unsustainable logging	12	35
16	This house would ban partial birth abortions	14	46
17	This house would build the Keystone XL pipeline	3	23
18	This house would embrace multiculturalism	25	64
19	This house would encourage the creation of private universities in the UK	5	8
20	This house would introduce year round schooling	4	13
21	This house would limit the right to bear arms	10	65
22	This house would only teach abstinence for sex education in schools	12	109
23	This house would re-engage with Myanmar	6	22
24	This house would reintroduce national service	6	17
25	This house would remove United States military bases from Japan	7	14
26	This house would use foreign aid funds to research and distribute software that allows bloggers and journalists in non-democratic countries to evade censorship and conceal their online activities	5	8

Table 5.9: Topic ids, topic titles, number of documents, and number of pieces of evidence in the ED-EMNLP-2015 dataset.

out-of-domain training data for the ED experiments.

Argument Mining Dataset

UKP Sentential AM This dataset consists of around 25k sentences distributed over eight topics. The sentences were extracted by first querying an external search engine with the topic and then selecting 50 documents archived on the Internet Archive⁶. Of these 50 documents per topic, Stab et al. (2018b) removed the boilerplate and all sentences that did not contain a verb or were less than three tokens long. The remaining sentences were then annotated by crowdworkers, meaning they contain either “evidence or reasoning that can be used to support or oppose a given topic” (Stab et al., 2018b), or are non-argumentative. Argumentative sentences are also annotated as either being in favour or against the topic.

5.2.2 Dataset Preparation

To be useable in our simulations, we needed to convert the datasets from collections of unannotated documents and lists of pieces of evidence to evidence-annotated documents. This means that we traced all pieces of evidence back to their original documents and labelled those sentences that are evidential in the documents.

ED-ACL-2014 and ED-EMNLP-2015 Before mapping the evidential sentences to the original document, we read the labelled pieces of evidence and segmented them into individual sentences. We found that some pieces of evidence did not map directly to the sentences in the source documents but only to parts of sentences.⁷ We therefore decided that if the token sequence labelled as evidence is part of a sentence, then that entire sentence is a piece of evidence. Listing B.2 shows the underlying method we used to conduct the mapping. The parameter `sentences` must be in the same order in which they appear in the original document. The `evidential_sentences` is the set of evidence that we segmented into sentences and filtered by length. The result `annotations` of the function is the sequence of labels for each sentence in the original document. We ignored all pieces of evidence shorter than three tokens to avoid having problems in mapping. For instance, we found that after sentence segmentation, some sentences consisted only of the final period, which then caused all sentences to be labelled as evidence. We used NLTK⁸ for pre-processing.

UKP Sentential AM Because the UKP Sentential AM dataset does not contain any arguments that are more than one sentence long, we did not need to use such a sophisticated method for mapping. Instead, we extended the code provided by Stab et al. (2018b) to label all occurrences of an argument in the same document. The originally provided data contains the source URL, the URL to the archived version, the stance, and the MD5 hash for the specific sentence. We modified this code so that in addition to adding the sentence to the provided data, it also saves a copy of the archived document with the labelled sentences. We did not distinguish between supporting or contradicting arguments.

⁶<https://web.archive.org/>

⁷We think that this is due to changes in the underlying Wikipedia article between the original annotation and the later saving of the file.

⁸<https://www.nltk.org>

5.2.3 Dataset Statistics

The conversion changed the statistical properties of the modified datasets. In the case of the ED-ACL-2014 and ED-EMNLP-2015 datasets, the segmentation of pieces of evidence into individual sentences increased the number of pieces of evidence; it also increased the class imbalance of the UKP Sentential AM dataset. Table 5.10 shows an overview of the ED and AM datasets after conversion.

	Documents	Sentences	Evidence
ED-ACL-2014			
train test	143	20649	1318
ED-EMNLP-2015			
train dev	170	28540	2300
train test	234	35877	2646
UKP Sentential AM			
Original	–	25492	11139
Converted	400	39577	11538 ⁹

Table 5.10: Statistics of the ED datasets after conversion to sentence-level annotated documents.

5.2.4 Models

We used three models in our evaluation of the different transfer learning approaches. One model, namely $\text{bilstm}_{\text{direct}}$, was trained exclusively on the in-domain data, while the $\text{bilstm}_{\text{fine}}$ model was first pre-trained on the out-of-domain data before continuing the training on the in-domain data; before continuing the training on the in-domain data; we refer to the latter model as $\text{bilstm}_{\text{pre}}$. For hyper-parameter tuning, we used the train-dev part of the ED-EMNLP-2015 dataset and the $\text{bilstm}_{\text{direct}}$ model. We did not conduct further hyper-parameter tuning for the $\text{bilstm}_{\text{fine}}$ model.

In our transfer training evaluation, we chose two different models. As our first transfer learning model, we used a bilstm which we pre-trained on the out-of-domain data before fine-tuning on the in-domain data. The architecture and hyper-parameters were the same as the $\text{bilstm}_{\text{direct}}$. As the direct transfer model, we selected the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) because it performed very well on both tasks and on the out-of-domain data. We conducted no hyper-parameter tuning on any of the transfer learning models.

$\text{bilstm}_{\text{direct}}$ For the direct training evaluation, we chose the $\text{bilstm}(100)$ model from the previous direct training evaluation. It uses 100-dimensional GloVe embeddings (Pennington, Socher, and Manning, 2014) as the input features and a dropout for each layer of 0.5. We trained this model for 10 epochs on the in-domain data.

$\text{bilstm}_{\text{fine}}$ The $\text{bilstm}_{\text{fine}}$ model’s architecture and hyper-parameters are identical to the $\text{bilstm}_{\text{direct}}$ model. Before fine-tuning the model on the in-domain data, we pre-trained it for five epochs on the out-of-domain training data; we refer to the resulting pre-trained model as $\text{bilstm}_{\text{pre}}$. For fine-tuning, we chose an approach similar to gradual unfreezing

⁹The number is different from the original dataset because of duplicated evidential sentences. There are 11, 128 unique pieces of evidence in the converted dataset.

(Howard and Ruder, 2018). This means that we first replaced the dense layer and trained it exclusively for five epochs with the other layers being frozen and, therefore, not receiving any update. Afterwards, we unfroze the rest of the network and trained the complete network for five additional epochs. We kept the same learning rate of 0.001 and dropout of 0.5. Figure 5.5 illustrates the replacement and continued training.

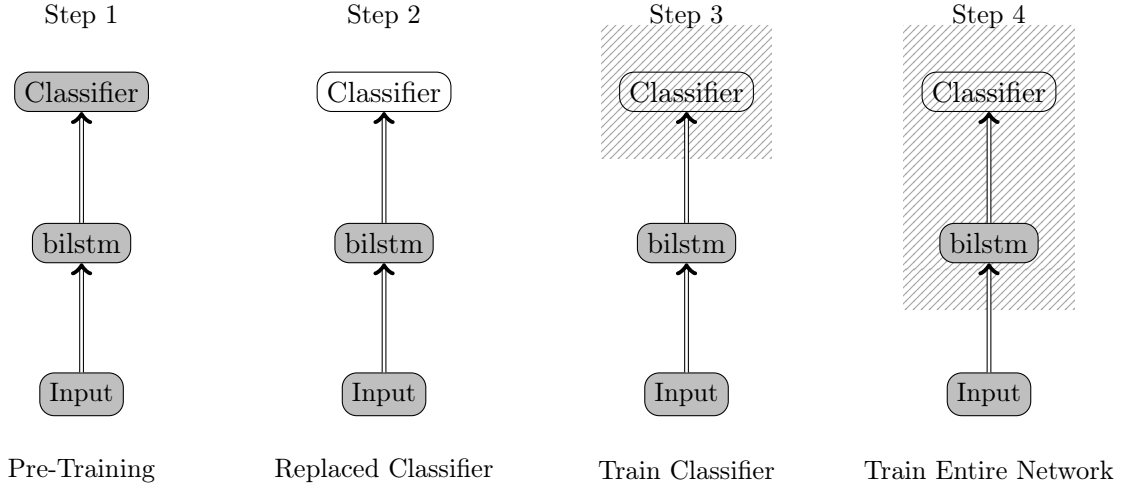


Figure 5.5: Illustration of the fine-tuning for the $\text{bilstm}_{\text{fine}}$ model. The layer without background colour is added in the second step, and the nodes with a patterned background are trained in this step.

BERT We selected the BERT base model and implementation from Huggingface¹⁰ for the direct transfer approach. We fine-tuned it for three epochs on the out-of-domain data. Because the out-of-domain data covers multiple topics, we used both the topic and candidate sentence as input. This means that the input to BERT is $\$TOPIC \langle SEP \rangle \$CANDIDATE_SENTENCE$ as one sequence of tokens. We decided not to train BERT on the in-domain data, because we think that the in-domain data was too small.

Table 5.11 shows the relation between the model and training domain.

	Training domain	
	Out-of-Domain	In-Domain
BERT	yes	no
$\text{bilstm}_{\text{pre}}$	yes	no
$\text{bilstm}_{\text{direct}}$	no	yes
$\text{bilstm}_{\text{fine}}$	yes	yes

Table 5.11: Model label depending on the training data.

Random The random baseline learns to predict evidence according to the class distribution. We used the same random baselines as in the direct training experiments.

5.2.5 Evaluation

Before evaluating the performance of the transfer learning approaches, we needed to ascertain the quality of the models on the out-of-domain data, or the testing data of their

¹⁰<https://github.com/huggingface/pytorch-pretrained-BERT>

respective training domain. To reduce the influence of the random initialisation of the models, we repeated the experiments with 10 different randomisation seeds $s \in \{0, 1, \dots, 9\}$.

Evaluation of Pre-trained Models in their Source Domain

To assess the pre-trained ED models, we evaluated them on the testing data of their source domain. This means that when pre-training for ED, the evaluation is conducted on the testing data of the ED-ACL-2018 dataset, and when pre-training for AM, the testing data was the test-split data of the training topics. We found that the `bilstmpre` model reached an accuracy of about 0.64, which is about 0.1 below the results presented by Shnarch et al. (2018). BERT, on the other hand, reached an accuracy of about 0.80, which is similar to the results published by Reimers et al. (2019). Table 5.12 shows the macro-averaged results for both: ED and AM tasks. In the case of ED, it also shows the accuracy for comparability with published literature.

For AM, we conducted the experiments in a leave-one-topic-out fashion. We trained the model for each left out topic on the training data of all but the left out topic and tested it on the testing data of the left out topic. The results of `bilstmpre` are comparable to the baseline established in Stab et al. (2018b), with a macro-averaged F1 score of around 61%. BERT reached a macro-averaged F1 score of almost 0.80, which sets a new state-of-the-art for this task in a two-label configuration.

	F1	P	R	Accuracy
ED-ACL-2018				
<code>bilstm_{pre}</code>	0.609	0.620	0.608	0.639
BERT	0.781	0.809	0.770	0.802
Reimers et al. (2019)	–	–	–	0.814
UKP Sentential AM				
<code>bilstm_{pre}</code>	0.624	0.647	0.632	–
BERT	0.795	0.800	0.800	–
Stab et al. (2018b)	0.666	–	–	–

Table 5.12: Results of the pre-trained models on the test data of their source domain. The results are macro-averaged for F1, precision, and recall.

Evaluation in the In-Domain

We conducted the transfer learning experiments in a leave-one-document-out fashion for each topic separately. This means that for each left out document, we took the remaining as training data, and trained or fine-tuned a model on it. For BERT, we conducted no additional training on the in-domain data and tested the already trained model on all documents for each topic. Table 5.13 shows the performance on the in-domain data of the transfer learning experiments.

Although all models beat the random baseline in evidence-F1 score, there is a large variation between the models on both ED datasets. Whereas BERT reached the highest precision on both ED datasets, its recall was generally lower than that of the in-domain trained models. For the ED-EMNLP-2015 dataset, the in-domain trained models outperformed BERT, except in precision. Furthermore, although all models performed better on the ED-EMNLP-2015 dataset than on the ED-ACL-2014 dataset, the models trained on the in-domain benefited more than BERT.

		F1	P	recall
ED-ACL-2014				
Evidence & No Evidence	Random	0.494 (0.000)	0.495 (0.001)	0.495 (0.000)
	bilstm _{direct}	0.509 (0.033)	0.514 (0.028)	0.526 (0.039)
	bilstm _{fine}	0.481 (0.043)	0.518 (0.018)	0.553 (0.047)
	BERT	0.540 (0.052)	0.590 (0.055)	0.538 (0.048)
Evidence only	Random	0.052 (0.001)	0.051 (0.000)	0.053 (0.002)
	bilstm _{direct}	0.117 (0.058)	0.091 (0.055)	0.183 (0.053)
	bilstm _{fine}	0.139 (0.064)	0.088 (0.045)	0.373 (0.118)
	BERT	0.118 (0.098)	0.238 (0.105)	0.094 (0.096)
ED-EMNLP-2015				
Evidence & No Evidence	Random	0.495 (0.000)	0.495 (0.001)	0.495 (0.000)
	bilstm _{direct}	0.572 (0.062)	0.566 (0.050)	0.613 (0.075)
	bilstm _{fine}	0.544 (0.063)	0.553 (0.046)	0.631 (0.089)
	BERT	0.550 (0.060)	0.596 (0.084)	0.558 (0.081)
Evidence only	Random	0.056 (0.000)	0.056 (0.003)	0.057 (0.001)
	bilstm _{direct}	0.225 (0.133)	0.176 (0.114)	0.340 (0.160)
	bilstm _{fine}	0.212 (0.132)	0.145 (0.101)	0.453 (0.212)
	BERT	0.143 (0.118)	0.251 (0.169)	0.143 (0.171)
UKP Sentential AM				
Evidence & No Evidence	bilstm _{fine}	0.681 (0.021)	0.698 (0.014)	0.739 (0.021)
	BERT	0.754 (0.016)	0.747 (0.015)	0.779 (0.015)
Evidence only	bilstm _{fine}	0.620 (0.027)	0.490 (0.034)	0.848 (0.015)
	BERT	0.676 (0.023)	0.599 (0.033)	0.780 (0.038)

Table 5.13: The results on the in-domain datasets are macro-averaged across all topics, with the standard deviations shown in parenthesis.

When analysing the performance of the models on the ED-ACL-2014 dataset in more detail, which is shown in table 5.14, we find that for all topics, at least one model outperformed the random baseline. The bilstm_{fine} most often reached the highest evidence-F1 score. More specifically, it reached the highest evidence-F1 score for half of all topics. For the two topics in which the bilstm_{direct} model reached the highest score, the bilstm_{fine} model still outperformed BERT.

Id	Docs	Evidence/Docs	Random	bilstm _{direct}	bilstm _{fine}	BERT
0	6	1.00	0.000	0.072	0.082	0.366
1	19	20.67	0.070	0.000	0.155	0.025
2	10	2.70	0.026	0.068	0.059	0.129
3	11	6.73	0.092	0.235	0.248	0.086
4	13	2.69	0.048	0.076	0.096	0.117
5	10	4.6	0.069	0.189	0.159	0.040
6	13	5.46	0.038	0.072	0.074	0.014
7	6	5.50	0.040	0.114	0.101	0.062
8	13	5.92	0.063	0.060	0.103	0.082
9	15	8.80	0.066	0.208	0.271	0.251
10	20	6.70	0.053	0.000	0.163	0.167
11	7	5.14	0.055	0.116	0.157	0.074

Table 5.14: Number of documents, averages pieces of evidence per document, and evidence-F1 scores of the random baseline, bilstm_{direct}, bilstm_{fine} models, and BERT for each topic on the ED-ACL-2014 dataset.

On the ED-EMNLP-2015 dataset, both the $\text{bilstm}_{\text{direct}}$ and $\text{bilstm}_{\text{fine}}$ models outperformed the random baseline on all topics, except for topic 19. BERT did not outperform the random baseline on topics 4, 9, 20, 23, 24, and 26. The $\text{bilstm}_{\text{direct}}$ model reached the highest evidence-F1 score for most topics (15 out of 26), followed by the $\text{bilstm}_{\text{fine}}$ model (6 out of 26) and BERT (6 out of 26), as shown in table 5.15.

Id	Docs	Evidence/Docs	Random	$\text{bilstm}_{\text{direct}}$	$\text{bilstm}_{\text{fine}}$	BERT
0	5	3.00	0.042	0.106	0.057	0.157
1	11	15.27	0.127	0.382	0.469	0.213
2	4	1.5	0.000	0.018	0.007	0.345
3	4	1.25	0.018	0.187	0.171	0.098
4	3	1.33	0.073	0.363	0.358	0.066
5	13	1.38	0.040	0.151	0.122	0.080
6	2	1.50	0.000	0.000	0.000	0.554
7	14	10.43	0.143	0.421	0.424	0.261
8	4	2.50	0.030	0.071	0.107	0.068
9	4	4.25	0.060	0.330	0.322	0.024
10	17	2.06	0.022	0.122	0.124	0.162
11	8	3.75	0.069	0.328	0.270	0.177
12	15	3.67	0.047	0.194	0.156	0.182
13	9	3.22	0.075	0.226	0.201	0.190
14	12	3.25	0.042	0.210	0.203	0.156
15	12	2.92	0.043	0.228	0.159	0.151
16	14	3.29	0.101	0.381	0.434	0.131
17	3	7.67	0.049	0.257	0.244	0.187
18	25	2.56	0.051	0.357	0.265	0.148
19	5	1.60	0.030	0.010	0.000	0.034
20	4	3.25	0.035	0.042	0.091	0.017
21	10	6.50	0.082	0.445	0.352	0.155
22	12	9.08	0.092	0.370	0.349	0.253
23	6	3.67	0.068	0.164	0.137	0.004
24	6	2.83	0.035	0.188	0.173	0.003
25	7	2.00	0.015	0.149	0.189	0.032
26	5	1.60	0.060	0.371	0.350	0.024

Table 5.15: Number of documents, averages pieces of evidence per document, and evidence-F1 scores of the random baseline, $\text{bilstm}_{\text{direct}}$, $\text{bilstm}_{\text{fine}}$ models, and BERT for each topic of the ED-EMNLP-2015 dataset.

For AM, BERT almost always reached a better result than the $\text{bilstm}_{\text{fine}}$, where its additional fine-tuning improved the recall by about 0.1 compared with the pre-training evaluation. The $\text{bilstm}_{\text{fine}}$ generally improved during the additional fine-tuning by 0.05 – 0.10, whereas BERT lost about 0.05 because of the different evaluation setting. This resulted in a reduction in performance by almost 0.1, from 0.171 to 0.073. Table 5.16 shows the performance of both models for each topic in the UKP Sentential AM dataset.

5.2.6 Discussion

ED-ACL-2014 We found no correlation between the number of files available to train or fine-tune a model and its performance. The $\text{bilstm}_{\text{direct}}$ model reached the highest score for topics 5 and 7, which contain 10 and 6 documents, respectively. The $\text{bilstm}_{\text{fine}}$ model reached the highest score for topics with over 10 documents. However, for topics 0, 2, 4, and 10, which contain 6, 10, 13, and 20 documents, the best-performing model is BERT. Which model performs best, therefore, does not depend on the number of available

Topic	bilstm _{fine}	BERT
Abortion	0.575 (0.006)	0.636
Cloning	0.658 (0.003)	0.701
Death Penalty	0.588 (0.006)	0.645
Gun Control	0.634 (0.008)	0.678
Marijuana Legalization	0.641 (0.004)	0.69
Minimum Wage	0.643 (0.006)	0.706
Nuclear Energy	0.618 (0.006)	0.685
School Uniforms	0.606 (0.008)	0.668

Table 5.16: Evidence-F1 scores of the bilstm_{fine} model and BERT for each topic of the UKP Sentential AM dataset.

training documents. When analysing the performance of the in-domain trained models, we found that the average number of pieces of evidence per document is a good indicator. For instance, although topic 9 and 10 have similar amounts of evidence, at 134 and 132, respectively, the performance of the trained models on topic 9 is considerably higher than on topic 10. We think that this is due to the lower number of documents over which the evidence is distributed for topic 9: 15 rather than 20. Similarly, topic 3 has 74 pieces of evidence distributed over 10 documents, and it reaches the second highest performance. On the lower end of performance, we found that topic 0 having an average of one piece of evidence per document and reached one of the lowest scores, together with topic 2 and topic 4 (2.7 pieces of evidence per document for either topic). One anomaly in this explanation is the low performance of the bilstm_{direct} model on topic 1. We conclude that the class balancing is not effective enough to counter the large imbalance found in the dataset.

ED-EMNLP-2015 As with the ED-ACL-2014 dataset, the number of documents per topic is not a reliable predictor for success of any method. Although most topics on which BERT reached the highest performance contain fewer than 6 documents, others contain fewer than 5 documents, and both in-domain trained models outperformed BERT in evidence-F1 score. Additionally, the number of documents is also no indicator regarding the choice of an in-domain trained model. Although it is reasonable to assume that few documents are beneficial for fine-tuning a pre-trained model and that large collections of documents are more suited for direct training, we found no evidence favouring it. Furthermore, although we observed the best performance only for topics with more than 10 documents, having many documents is not a sufficient condition for higher performance. Similar to the previous dataset, the average number of pieces of evidence is a good indicator whether the characteristics of evidence for a topic can be learned. Topic 26 seems to perform surprisingly well, considering that it only contains five documents and eight pieces of evidence. Upon further inspection, we found that this is due to an error in the sentence segmentation, turning a reference into a separate sentence. The sequence “*/ref/*.” was therefore labelled evidence in multiple documents and is easy to learn. Although the sentence segmentation error also occurred in other topics, they not found to contain this sequence in their list of evidence.

Performance increase between ED-ACL-2014 and ED-EMNLP-2015 datasets

We found that all methods varied depending on the dataset and that their performance improved with newer datasets. To test this hypothesis in more detail, we compared the performance of all three models on the 12 topics of the ED-ACL-2014 with the performance on the same topics of the ED-EMNLP-2015 dataset. Although the topics are the same,

the datasets are not completely identical. Some hypotheses were added, others removed, and some evidence has been added. Table 5.17 shows that the in-domain trained models improved in performance on ED-EMNLP-2015 dataset over the ED-ACL-2014 dataset. Both in-domain trained models improved by about 0.1 evidence-F1 score. However, we found that BERT’s performance dropped by almost 0.01 evidence-F1 score.

	F1	P	R
bilstm _{direct}	0.243 [+0.126]	0.195 [+0.104]	0.345 [+0.162]
bilstm _{fine}	0.233 [+0.094]	0.155 [+0.067]	0.509 [+0.136]
BERT	0.110 [−0.008]	0.290 [+0.052]	0.073 [−0.021]

Table 5.17: Results on the 12 topics of the ED-EMNLP-2015, which are also present in the ED-ACL-2014 dataset. The scores are on the evidence class and macro-averages across the topics. The values in brackets are the difference from the ED-ACL-2014 dataset.

Poor Performance of BERT Although BERT performed very well when tested on the ED-ACL-2018 dataset, the quality of its predictions dropped dramatically when tested on the ED-ACL-2014 or ED-EMNLP-2015 datasets. We formulated two hypotheses that would explain the drop and then evaluated them.

- (1) The wording of the topic labels varies too much between the ED-ACL-2018 and the ED-EMNLP-2015 datasets.
- (2) The segmentation into individual sentences of pieces of evidence spanning multiple sentences causes the dramatic drop in recall between the ED-ACL-2018 and the ED-EMNLP-2015 datasets.

The wording of the topic labels in the ED-ACL-2018 dataset is a controversial statement, such as “*We should ban gambling*”. In the ED-EMNLP-2015 dataset, the wording is that of a debate motion, or “*This house would ban gambling*”. We evaluated this effect of adapting the topic label of three topics that appeared in the training data of the ED-ACL-2018 dataset and the ED-EMNLP-2015 dataset. We took the data from the ED-EMNLP-2015 dataset but revised the topic label to be identical to the one used in the ED-ACL-2018 dataset. Table 5.18 shows that using the wording of the topic label from the training

	F1	P	R
in-domain topic label	0.077	0.213	0.050
out-of-domain topic label	0.087	0.262	0.060

Table 5.18: The results show only the evidence class and are macro-averaged across the three selected topics.

data increased the performance by 0.01 evidence-F1 score. Therefore, we conclude that the wording of the topic label had only minor effects. To test our second hypothesis, we tested BERT on all pieces of evidence on the three previously selected topics. We found that not segmenting the pieces of evidence into individual sentences led to an increase in recall by 0.04 to 0.098. We therefore conclude that the segmentation of pieces of evidence into individual sentences is also not the cause of the poor performance of the out-of-domain trained BERT. This leads us to conclude that the different annotation schemes in combination with the considerably larger number of annotators led to too large a domain shift between the out-of- and in-domain datasets. The ED-ACL-2018 dataset might be more consistent, because all sentences were labelled by ten annotators. This means that if

only a few annotators found that a sentence is a piece of evidence, this sentence would not be labelled as such. However, this individual position is more present in the ED-ACL-2014 and ED-ACL-2018 datasets because of the different annotation process.

Upper bound for out-of-domain trained models We attempted to ascertain the upper bound by manually labelling the documents for two topics. We selected the topics with the highest and lowest scores for BERT, i.e. topic 6 and topic 24. We then selected two documents for each topic for manual annotation and calculated the evidence-F1 score. For topic 6, we found no agreement with the evidence provided by the dataset. For topic 24, we reached an evidence-F1 score of 0.190 with a recall of 0.500 and precision of 0.118. We conclude that for a better evaluation of the upper bound, the annotation guidelines are required.

5.3 Interactive Learning

Training an ED model interactively means that a user interacts with a system that is learning to detect evidence in textual sources. The interaction can take many different forms, ranging from providing a score that indicates how well the system performs to generating data for supervised ML techniques. In this thesis, we treat ED as a classification task in which the user generates data that can be used in supervised ML. In this situation, an important question is how much training data are necessary for the ED model to be beneficial to the user. Many approaches that evaluate how much a trained model depends on the size of the training data use random sub-sampling to reduce the training data (e.g. Schulz et al., 2018). However, when the training data uses different sources, it is possible that each source only covers a subset of aspects or even a single one. When looking for evidence related to nuclear energy, an article on historical nuclear catastrophes will not likely contain evidence regarding the cost of nuclear waste management. Using random sub-sampling to evaluate the necessary amount of training data does not take this distribution into account. If the training data are reduced to half their original size and each document contains multiple pieces of evidence regarding each aspect, then the training data still covers many aspects. This means that a smaller dataset could still cover the feature space well enough to be used to train a well performing model. However, when we remove multiple aspects from the training data, the performance will be affected much more significantly. This is especially important when considering that real users do not start with a predefined set of hypotheses but instead develop them over time, as demonstrated in chapter 4.

5.3.1 User and System Simulation

We simulate users whose approach to finding evidence mirrors our findings from chapter 4. First, the majority of users worked on a single document at-a-time and did not return to a previously read document. Second, most users worked in the same order in which the documents were presented in EDoHa. This means they worked through the documents in alphabetical order.

Therefore, our simulated user reads the documents \mathcal{D} one-by-one in alphabetical order. While reading a document $d_t \in \mathcal{D}$, they label evidential sentences as evidence. Once they have finished reading a document d_1 , its sentences and labels are added to the training data. The system then trains an ED model m_1 on the currently available training data. When the user opens the next document d_2 , the system first uses the most recently trained ED model m_1 to predict whether each sentence in the document is evidential or not. The

user then corrects these predictions by fixing the label of incorrectly predicted pieces of evidence and labelling the sentences that the model missed. Afterwards, the sentences and labels are also added to the increasing amount of training data for the next model m_2 . This continues until the user opens the last document d_T , which contains predictions from the most recent model m_{T-1} . Figure 5.6 illustrates the simulated user and system with the growing amount of training data.

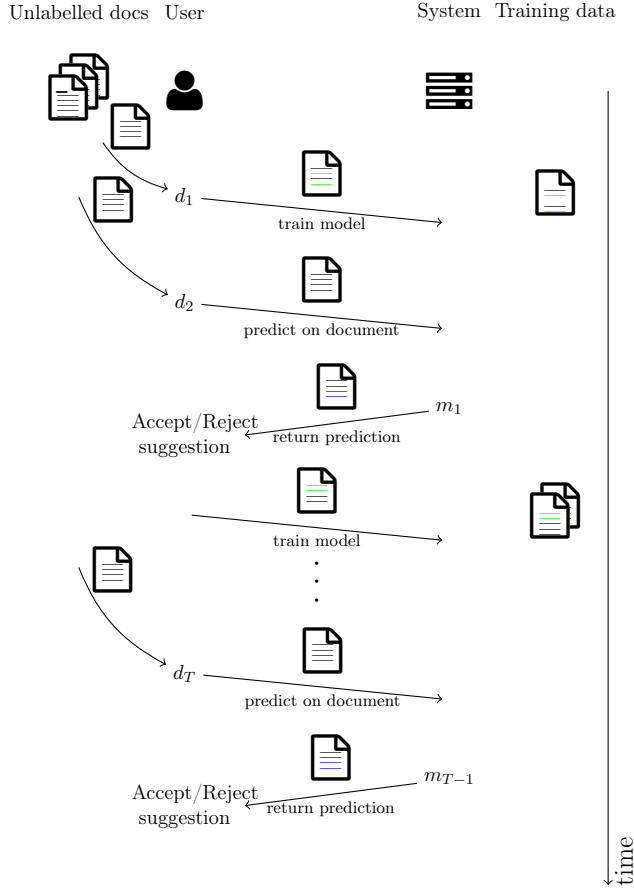


Figure 5.6: Sequence diagram of the simulated user interactively creating training data, training ED models, and correcting the predictions. The user picks an un-annotated document and labels the evidential sentences. After processing the document, it is added to the training data for a newly trained model. Afterwards, the user picks the next document, which contains suggestions from the model.

5.3.2 Measurement of User Effort

To evaluate the quality of the interactively trained models, we also need to measure the amount of work required by the user to correct the predictions. Using the evidence-F1 score is not beneficial in this case, because it calculates the harmonic average across both the precision and recall. We found F1 scores to behave counter-intuitively as a measure of effort, because when the quality increases, the effort should decrease. Instead, we focus on the number of clicks a user must make to correct the model's prediction.

Using click counts does not allow for a comparison across different documents because the number of sentences varies, along with the ratio of evidential and non-evidential sentences. For instance, suppose the first document consists of 100 sentences and the second

consists of 200; then, the click count it took to correct the same ED model is not comparable. If the model makes an error on every 50th sentence, this means that the first document contains two errors and the second four. Instead of click counts, we use the sum of the *false discovery rate* and *false omission rate* to define the *error-rate* E . The false discovery rate $fp/tp+fp$ is the rate of sentences that are incorrectly classified as evidence, and the false omission rate $fn/tp+fn$ is the rate of pieces of evidence missing from the classification. Formally defined, the error-rate

$$E = \frac{fp}{tp + fp} + \frac{fn}{tp + fn} = (1 - P) + (1 - R), \quad (5.1)$$

where P is the precision and R being the recall on the evidence class. Using the sum of the precision and recall rather than the average has two effects. First, if a model labelled a non-evidential sentence as evidence, it would take the user two clicks to correct it. The first one to de-select the incorrectly suggested piece of evidence and the second one to select the correct piece of evidence. Second, the error-rate has a value range from 0 to 2, with all values above 1 meaning that it would take the user less work to label all evidential sentences themselves rather than correcting the predictions from the ED model. We can calculate the error-rate based on the difference between the predicted pieces of evidence and the actual pieces of evidence present in the document.

5.3.3 Evaluation

In our evaluation of interactive learning, we conducted the experiments in a leave-one-document-out fashion. For each left out document, we first trained the interactively trained models on the first training document and tested it on the left out one. Then, we trained the model on the first two documents and tested it again on the left out document. This helped reduce the noise in the results that would appear because of the differences between the test document after one, two, or more training documents.

We determined the number of training documents required to outperform the baseline as

$$\mu = \begin{cases} \min \{t \mid \forall i \in \{t, \dots, T\} : \mathcal{R}_i < \mathcal{B}\} & , \exists t : \mathcal{R}_t < \mathcal{B} \\ |T| & , \text{otherwise,} \end{cases} \quad (5.2)$$

where \mathcal{R} is the sequence of error-rates over time $t \in \{1, \dots, T\}$ and \mathcal{B} is the average error-rate of the baseline. We then averaged the minimum number of training documents across the 10 different randomisation seeds.

Table 5.19 shows that for the ED-ACL-2014 dataset, the `bilstmfine` model almost always required less work for the user to correct the predictions and that it outperformed BERT often already after the first document. Of the 12 topics in the dataset, BERT only twice performed better than the `bilstmfine`. Both of these topics contained more documents than were required for the other topics for the `bilstmfine` model to outperform BERT.

On the ED-EMNLP-2015 dataset, both the interactively trained and fine-tuned models often required less work to correct and generally only required one training document to do so. Of the topics in which BERT performed best, neither contained more than five documents, and topic 6 only contained two. For three topics, the `bilstmdirect` model performed better than either BERT and `bilstmfine`. However, the difference to the `bilstmfine` model is much smaller than to BERT.

In only two cases would the user have to work less to correct the model's predictions than to label the sentences themselves. These cases are the `bilstmfine` model for topic 1 and BERT for topic 6; in the case of topics 7 and 21, the `bilstmfine` model almost reached this

Id	Docs	bilstm _{fine}		BERT
		μ	E	E
0	6	6.000	1.761	1.230
1	19	3.800	1.526	1.677
2	10	10.000	1.752	1.617
3	11	1.000	1.288	1.742
4	13	6.300	1.655	1.737
5	10	5.200	1.612	1.772
6	13	1.000	1.535	1.898
7	6	1.300	1.665	1.830
8	13	5.500	1.525	1.681
9	15	10.200	1.307	1.426
10	20	6.200	1.397	1.560
11	7	1.000	1.445	1.846

Table 5.19: Number of documents and minimum number of training documents μ to reach a smaller error-rate E than BERT for the bilstm_{fine} model for each topic on the ED-ACL-2014 dataset.

level of quality. For many topics, it took only one training document for the bilstm_{direct} and bilstm_{fine} models to outperform BERT.

Another observation is that the performance in evidence-F1 score does not mirror the performance in error-rate E . Whereas in evidence-F1 score the direct training approach most often reached the best performance, when measuring the work required by a user to correct the predictions, the fine-tuning approach most often performed best.

5.3.4 Discussion

We found that all error-rates, except two topics in the ED-EMNLP-2015 dataset, were above one. This means that except for the two topics where the error-rate was below one, the user would have to do more work to correct the suggestions the ED models made than to label all evidential sentences themselves.

For the two topics in which the error-rate reached a value below one, namely topic 1 and topic 6, we found that for topic 1, the bilstm_{fine} model reached an error-rate below one, and for topic 6, it was BERT. Upon closer inspection, we found that the bilstm_{fine} model reached an error-rate close to 1, after about six training documents and from then alternated slightly around this error-rate. In the case of topic 6, we found that BERT significantly outperformed both the bilstm_{direct} and the bilstm_{fine} models, because neither interactively trained model learned to predict evidence.

For the topics 7 and 21, the bilstm_{fine} model reached an error-rate close to 1, with 1.011 and 1.019, respectively. Therefore, we investigated whether the bilstm_{fine} model improved consistently with an increase in data so that we could conclude if more training data would have led to further improvement. In the case of topic 7, we found that the bilstm_{fine} model reached its lowest error-rate after about 5 training documents, after which it reached a saddle point with another improvement after 12 training documents. A similar pattern can be found for topic 21. The error-rate reached its lowest value after four training documents, after which the error-rate slightly increased with a continuing decrease after seven to nine training documents. We think that in the case of topic 7, it would require a large amount of additional training documents, because the error-rate only decreased slightly when using all available training documents. For topic 21, on the other hand, the error-rate decreased throughout multiple succeeding increases in training documents. Only a few additional documents would be required until its predictive quality would reach a level

Id	Docs	bilstm _{direct}		bilstm _{fine}		BERT
		μ	E	μ	E	E
0	5	5.000	1.739	5.000	1.837	1.586
1	11	1.000	1.194	1.000	0.932	1.373
2	4	4.000	1.932	4.000	1.981	1.226
3	4	1.000	1.432	1.200	1.474	1.793
4	3	1.000	1.235	1.000	1.247	1.799
5	13	1.000	1.643	1.000	1.591	1.829
6	2	2.000	2.000	2.000	2.000	0.723
7	14	1.000	1.123	1.000	1.011	1.472
8	4	4.000	1.805	4.000	1.592	1.289
9	4	1.000	1.244	1.000	1.182	1.607
10	17	17.000	1.592	5.000	1.268	1.385
11	8	1.400	1.307	1.000	1.329	1.636
12	15	14.000	1.554	4.800	1.433	1.543
13	9	8.800	1.486	6.500	1.405	1.416
14	12	2.200	1.484	1.200	1.297	1.683
15	12	3.100	1.500	1.000	1.366	1.643
16	14	1.000	1.213	1.000	1.049	1.724
17	3	2.300	1.449	1.700	1.401	1.490
18	25	1.000	1.190	1.000	1.152	1.517
19	5	4.700	1.960	5.000	2.000	1.903
20	4	2.600	1.862	2.200	1.688	1.949
21	10	1.000	1.039	1.000	1.019	1.606
22	12	1.100	1.240	1.000	1.147	1.431
23	6	3.000	1.585	3.000	1.478	1.990
24	6	1.000	1.456	1.000	1.350	1.965
25	7	2.700	1.590	3.800	1.461	1.923
26	5	1.000	1.217	1.000	1.151	1.871

Table 5.20: Number of documents and minimum number of training documents μ to reach a smaller error-rate E than BERT for the bilstm_{direct} and bilstm_{fine} models for each topic on the ED-EMNLP-2015 dataset.

at which their correction requires less effort from the user than labelling each evidential sentence manually.

5.4 Chapter Conclusion

In this chapter, we investigated different methods to support users in finding evidence. We evaluated direct training approaches, as well as different kinds of transfer learning for ED, and simulated users to realistically determine the amount of in-domain data necessary to outperform a direct transfer approach with a well generalising model, namely BERT, which had no access to the in-domain data.

To simulate our users, we first mapped the annotations of three datasets back to the original documents. This allowed us to simulate the behaviour of a user who processes one document after another. The user would read the first document and label all evidential sentences. Then, once the user is done with this document, its sentences and labels are added to the training data and an ED model is trained on it. When the user opens the second document, it already contains suggestions made by the previously trained model.

This allows us to answer our sub-questions.

- ① *Does a user-specific ED model trained on small, in-domain data outperform a well*

generalising, user-independent one trained on out-of-domain data?

We found that having access to the in-domain data is oftentimes better than using a state-of-the-art model trained on out-of-domain data. Additionally, we found that the number of available documents is not a good indicator for which approach works best. For topics with few documents, we found each approach to perform best occasionally, and for topics with many documents, we found the same. However, we found the average number of pieces of evidence per document to be a good indicator and that calculating the class weights to address the class imbalance was insufficient to counter the effects of class imbalance.

- ② *Can we use user-independent, out-of-domain data for pre-training to reduce the necessary amount of user-specific, in-domain training data?*

We found that pre-training a bilstm on out-of-domain data before fine-tuning it on the in-domain data is also better than training it directly on the in-domain data. The direct transfer approach with BERT only performed better if the domain shift is small, i.e. a different topic within the same dataset, rather than a different dataset.

- ③ *How much training data does a user have to generate so that the interactively trained models can outperform a well generalising, user-independent model?*

We found that the amount of in-domain training to outperform BERT is generally one or a few documents.

- ④ *How close to the user-specific data must the training data of the user-independent model be, until it performs equally well as the user-specific one?*

Finally, we found that BERT performs better when the domain shift is small. Still, fine-tuning a simpler model with in-domain data reduces the gap.

We conclude that if the domain shift is large, i.e. two different datasets with potential differences in annotation guidelines, it is better to interactively fine-tune a simple model, such as a bilstm, than to fine-tune BERT on out-of-domain data. If the domain shift is small, i.e. the same dataset but different topics, then an interactively trained, small ED model performs only slightly worse than BERT, which was fine-tuned on out-of-domain data. This is even more important when computational resources are scarce, e.g. when no access to a GPU is possible or only a little amount of memory is available. For instance, predicting the label of 5000 sentences takes less than 20 seconds with a bilstm(100) on a notebook equipped with a Core i7-8550U CPU and 16 GiB of memory, whereas BERT using bert-as-service¹¹ takes about 200 seconds on the same machine.

¹¹<https://github.com/hanxiao/bert-as-service>

Chapter 6

Machine Learning for Evidence Linking

Linking pieces of evidence to hypotheses is the second task in which we aim to support the users. This would allow a user to spend less time in finding a previously labelled piece of evidence in a large list when trying to link it to a hypothesis. Instead, they could either accept or reject the suggestions, which takes less time than searching. In this thesis, we define EL as follows:

Definition 2. Let e be a piece of evidence, h a hypothesis, and f a function which predicts the probability of the piece of evidence being linked to the hypothesis. The goal in EL is to learn the function.

$$f(e, h) = \begin{cases} 1 & , \text{if } e \text{ is linked to } h \\ 0 & , \text{otherwise} \end{cases} \quad (6.1)$$

Supporting users in this task can be addressed in many different ways. Much like with ED, we can use large external sources of data to build a generic, user-independent model; or we can use different kinds of transfer learning to interactively train a user-specific model. Therefore, in addressing our third research question, (How well do machine learning-based methods work for EL?), we formulate the following sub-questions:

- ① Can we train a user-dependent EL model directly from user-created data?
- ② Can we use user-independent data for pre-training to reduce the necessary amount of user-specific training data?
- ③ Does a user-specific EL model outperform a generic one that has been trained on out-of-domain data?
- ④ What is the relation between hyper-parameters and individual users?

We address these questions by comparing three different models, each of which is trained in a different way. First, we look at a generic model trained exclusively on external, out-of-domain data. Second, we examine a transfer learning model, which is pre-trained on the same external, out-of-domain data as the previous one but fine-tuned on the user-specific data. Third, we look at a directly trained model that is trained exclusively on user-specific

data. These models are also compared against a random baseline. We also investigate how the hyper-parameters shared across users affect the performance.

We conducted the experiments on the nuclear energy and forest dieback datasets created in chapter 4. Still, we must first update the datasets so that they contain links between hypotheses and pieces of evidence, as well as non-links. A non-link is a hypothesis and a piece of evidence not supporting or contradicting the hypothesis. We must also consider that the amount of data vary greatly between the individual users, hence leading to additional challenges when conducting any hyper-parameter tuning.

6.1 Data

We selected the *nuclear energy* and *forest dieback* datasets as the in-domain data and the ED-ACL-2018 dataset as the out-of-domain data. Before conducting any experiment, we need to convert the user-created links into classification datasets. This means that we need to create negative data, or *non-links*, for training. We performed this conversion for the in-domain and out-of-domain data.

Data preparation We created our classification datasets by randomly sampling unrelated pieces of evidence for each hypothesis to create *non-links*. In this way, we made sure that our random pair is not a user-created link. To make sure that the same non-link did not appear in the training and testing splits, we treated each created non-link like a user-created link afterwards. Listing B.3 shows our method in more detail.

In-domain data Both the *nuclear energy* and *forest dieback* datasets contain links between pieces of evidence and hypotheses for each user. The number of links between pieces of evidence and hypotheses in the nuclear energy dataset varies between 12 and 259. In the forest dieback dataset, the number of links between pieces of evidence and hypotheses varies between 13 and 183. User12 in the nuclear energy dataset and User10 in the forest dieback dataset each reached a point after which no additional non-links could be created because all pieces of evidence were either already linked to a hypothesis or had been added as non-links. Table 6.1 shows the statistics of the resulting datasets.

Out-of-domain data Although the ED-ACL-2018 dataset is provided in English and our in-domain data are German, we were still able to use these data by translating it automatically from English to German via an external machine translation engine.¹ Afterwards, we created as many non-links as there are links between the topics and pieces of evidence. We did not filter out any duplicated non-links because it is not required because of the corpus being split into a training and testing dataset without overlap between the training and testing data. Table 6.2 shows the dataset statistics.

6.2 Models

To investigate which ML approach works best for EL, we built different models and compared them against a random baseline. The models, we evaluated are a Siamese bilstm with attention trained directly on the user-specific data and a similar network that is first pre-trained on external, out-of-domain data and then fine-tuned on the user-specific data. The last model we evaluated is BERT which we fine-tuned on the external, out-of-domain data.

¹<https://translate.google.com/>

Nuclear Energy					
	Links	Non-links			
User0	259	259			
User1	12	12			
User4	27	27			
User5	63	63			
User7	70	70			
User12	29	22			
User13	30	30			
User14	23	23			
User15	32	32			
User16	28	28			
User17	61	61			
User18	44	44			
User19	56	56			
User20	21	21			

Forest Dieback		
	Links	Non-links
User0	183	183
User1	45	45
User2	69	69
User3	61	61
User4	19	19
User5	26	26
User6	26	26
User7	46	46
User9	13	13
User10	29	18
User11	32	32

Table 6.1: Number of links and non-links per user of the nuclear energy (left) and forest dieback (right) datasets. For User12 in the nuclear energy and User10 in the forest dieback dataset, we could not create more unique non-links.

Split	Topics	Links	Non-links
Train	83	1499	1499
Test	35	683	683

Table 6.2: Statistics of the ED-ACL-2018 dataset with the additional non-links.

Random We implemented a random baseline, because not all users had a completely balanced dataset. The random baseline is similar to the one used in chapter 5; this means it learns the class distribution from the training data, and in prediction, it samples from this distribution.

bilstmAtt We selected a Siamese bilstm with two input bilstms: one for the candidate sentence and one for the hypothesis. The hypothesis bilstm encodes its input into a single vector representation, whereas the sentence bilstm encodes a contextual token representation.² We then calculate the attention similar to the token-level attention developed by Yang et al. (2016) between the encoded hypothesis and each token of the contextual sentence representation. The attention-weighted representation is then fed into a dense layer for classification. Between the attention and dense layers, we included a dropout layer for regularisation. Figure 6.2 shows the architecture of this network.

bilstmAtt_{fine} We first pre-trained the bilstmAtt model on the ED-ACL-2018 dataset and then continued to fine-tune it on the user-specific data. We first replaced the classification layer and then trained the new classification layer for ie epochs before doing so for the entire network for e epochs. The hyper-parameters ie and e were tuned on a different dataset of a similar size.

BERT We used a multilingual BERT (Devlin et al., 2018) by fine-tuning it on the training data of the ED-ACL-2018 dataset, with additional non-links translated to German.

²We decided against calculating the contextual token representation for the hypotheses, because many hypotheses consist of few or even a single word.

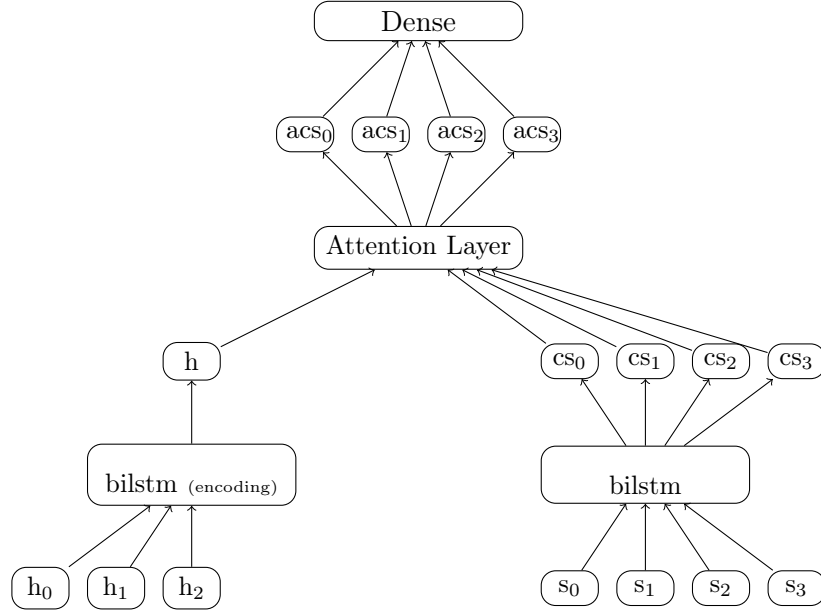


Figure 6.1: Architecture of the bilstmAtt network. The individual tokens of the hypothesis (h_0, h_1, h_2) are encoded into the individual hypothesis vector (h). The tokens of the sentence (s_0, s_1, s_2, s_3) are encoded into contextual token embeddings (cs_0, cs_1, cs_2, cs_3) and then weighed with hypothesis attention. The attention-weighted contextual token embeddings ($acs_0, acs_1, acs_2, acs_3$) are then fed into a dense layer for classification.

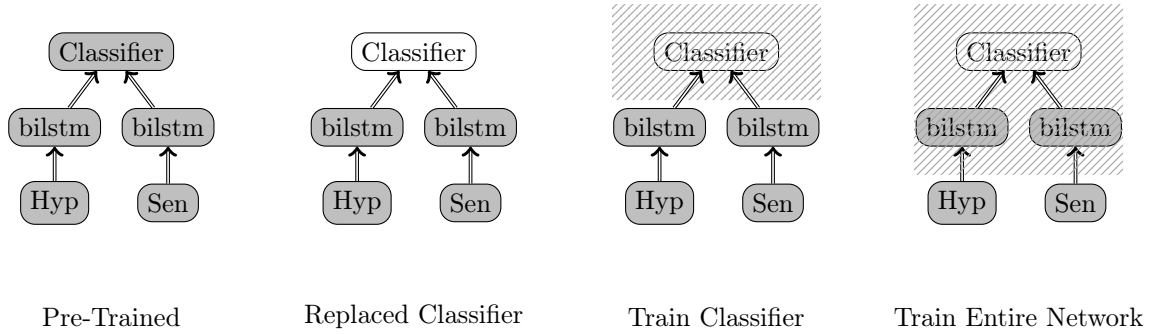


Figure 6.2: Transfer learning approach for the bilstmAtt_{fine}. The grey nodes are pre-trained, and the node without background colour is the newly added classifier. The nodes with the patterned background are trained or fine-tuned in the current step.

We fed the hypothesis and piece of evidence as a single sequence separated by the `<SEP>` token. As the basis for our fine-tuning, we chose the `bert-base-multilingual-cased` provided by Huggingface³. We fine-tuned it for three epochs on the training data of the ED-ACL-2018 dataset.

6.3 Hyper-Parameter Tuning

We decided against conducting individual hyper-parameter optimisations for each user because doing so would further decrease the already small size of the user-specific data. Instead, we conducted the hyper-parameter optimisation of analogous data of a similar

³https://huggingface.co/transformers/pretrained_models.html

size. We created groups of users with a similar amount of data to reduce the number of separate hyper-parameter optimisations. This means that we created groups of users with a similar number of links and non-links and conducted a separate hyper-parameter optimisation for each of these groups. We created four groups along the \log_2 scale of the number of links and non-links each user created in either dataset. This means we did not create separate groups for the nuclear energy and forest dieback datasets. Table 6.3 shows the groups, their boundaries, and the users who are part of this group.

Group	Min	Max	Users	
			Nuclear Energy	Forest Dieback
1	24	51	User1, User12, User14, User20	User4, User9, User10
2	52	111	User4, User13, User15, User16, User18	User1, User5, User6, User7, User11
3	112	240	User5, User7, User17, User19	User2, User3
4	241	517	User0	User0

Table 6.3: Group id, minimum and maximum number of links and non-links, and users for each group of both the nuclear energy and the forest dieback datasets.

6.3.1 bilstmAtt

We selected the ED-ACL-2018 dataset for the hyper-parameter optimisation of the direct training approach and down sampled the existing training data to a similar size for each user group. We then used a three-fold cross-validation on this analogous data to evaluate the hyper-parameter settings. We created a subset of the original training data for each group, of which table 6.4 shows the overall size and distribution between the links and non-links.

	Size	Links	Non-Links
Group 1	50	27	23
Group 2	100	59	41
Group 3	200	107	93
Group 4	400	207	193

Table 6.4: Statistics on the down-sampled datasets for the hyper-parameter tuning.

We decided to not optimise parameters that build the architecture of a network, such as the number of layers and nodes. We based this decision on the fact that if users with different amounts of links and non-links would use different networks, it would require changing the architecture while a user is working, hence creating more data. Instead, we focus on the learning rate, the number of epochs, and the batch size.

For hyper-parameter optimisation, we first selected four different values for each hyper-parameter. Table 6.5 shows these four values for each parameter. We then performed a grid search on all possible combinations.

Hyper-parameter	Possible Values			
Epochs	5	10	20	40
Learning rate	0.0001	0.0005	0.001	0.002
Batch size	1	2	4	16

Table 6.5: Possible values for the hyper-parameter optimisation.

6.3.2 bilstmAtt_{fine}

In the direct training experiments, we created a down-sampled version of an external dataset for hyper-parameter tuning. To optimise the transfer learning experiments, we cannot use these datasets, because we already used all of the training data for pre-training, leaving no development data. Therefore, we used an approach similar to co-training for hyper-parameter tuning. We used data from another user in the same group but from the other dataset. As the representative of a group, we selected the user with the largest number of links between pieces of evidence and hypotheses. For instance, to optimise the hyper-parameters of User5 in the nuclear energy dataset, we evaluated multiple hyper-parameter settings on the data from User2 in the forest dieback dataset. Table 6.6 shows which user represents their group for the co-hyper-parameter optimisation.

	Nuclear Energy	Forest Dieback
Group1	User12	User10
Group2	User18	User7
Group3	User7	User2
Group4	User0	User0

Table 6.6: Representatives of each user group for the nuclear energy and forest dieback datasets.

In contrast to the hyper-parameter tuning in the direct training experiments, we chose to use a random search for hyper-parameter tuning. We defined a set of values for each parameter and randomly sampled 60 of these values. We repeated this for each hyper-parameter, hence creating 60 random sets of hyper-parameter values. Table 6.7 shows the possible values for each hyper-parameter in the transfer-learning hyper-parameter optimisation.

Hyper-parameter	Possible Values				
<i>ie</i>	0	1	2	5	10
<i>e</i>	0	1	2	5	10
<i>lr</i>	0.0001	0.0005	0.001	0.002	
<i>ilr</i>	0.0001	0.0005	0.001	0.002	
batch size	1	2	4	16	

Table 6.7: The different possible values for the number of epochs to train and the classification layer alone *ilr*, the number of epochs to train the entire network *e*, the learning rate in the initial training phase *ilr*, the learning rate when training the entire network *lr*, and the batch size.

6.4 Results

In chapter 4, we analysed how users approach the task of EL and found that the majority of users worked on multiple hypotheses at-a-time. Therefore, we simulated our users with a cross-validation, rather than in a leave-one-hypothesis-out fashion.

Evaluation of Pre-trained Models

We evaluated the pre-trained models on the testing data of the ED-ACL-2018 dataset to ascertain their quality. BERT reached a macro-F1 score of over 0.95, whereas the $\text{bilstmAtt}_{\text{pre}}$ reached a macro-F1 score of 0.637. On the link class alone, BERT reached an F1 score of 0.946 and $\text{bilstmAtt}_{\text{pre}}$ of 0.692. Table 6.8 shows the performance scores of the pre-trained BERT and $\text{bilstmAtt}_{\text{pre}}$ models and a random baseline.

		F1	P	R
BERT	Macro	0.947	0.949	0.947
	Link	0.946	0.977	0.916
$\text{bilstmAtt}_{\text{pre}}$	Macro	0.637	0.666	0.647
	Link	0.692	0.616	0.796
Random	Macro	0.507	0.507	0.507
	Link	0.501	0.507	0.496

Table 6.8: Evaluation of the pre-trained EL models on the test data of the ED-ACL-2018 dataset. The results on the link and non-link classes are macro-averaged.

Hyper-Parameter Tuning

Direct Training In the hyper-parameter tuning for the bilstmAtt network we found that the performance of the models varied greatly between different settings and also groups. No hyper-parameter configuration reached a macro-F1 score of ≥ 0.570 , but several configurations performed below 0.5, which is about the quality of a random prediction for a balanced binary dataset. It is notable that the group with the lowest amount of data (Group1) reached the highest macro-F1 score with 0.562, whereas the group with the largest amount of data (Group4) reached the lowest with 0.515. All results can be found in the appendix in table C.1.

Fine-tuning for nuclear energy dataset The performance of the fine-tuned networks varies even more between the different hyper-parameter configuration than the direct training configurations. Similarly, no configuration reached a macro-F1 score of 0.57. The best-performing configuration reached a macro-F1 score of 0.52. All randomly sampled hyper-parameter configurations and the results can be found in the appendix in table C.2.

Fine-tuning for forest dieback dataset On the forest dieback dataset we can see that similar to the nuclear energy dataset, many hyper-parameter configurations performed below random chance. However, in contrast to the nuclear energy dataset, two groups have hyper-parameter configurations in which the macro-F1 score is above 0.5: Group2 and Group4 —and Group1 has a configuration in which the macro-F1 score is 0.5. All randomly sampled hyper-parameter configurations and the results can be found in the appendix in table C.3.

Simulated Users

In the user simulations, we found that the out-of-domain trained BERT performed best across all users in macro-averaged scores across both classes (link and no-link). On the nuclear energy dataset, it reached a macro-F1 score of 0.586, and on the forest dieback dataset, it reached a macro-F1 score of 0.550. On the link class, we found that BERT

reached the highest precision, but in the recall and F1 score, the random baseline performed the best. Although neither the `bilstmAtt` nor the `bilstmAttfine` networks outperformed the random baseline, the pre-training of the `bilstmAttfine` network improved the macro-F1 score by 0.12 and the link-F1 score by 0.062. On the forest dieback dataset, the BERT model again reached the highest macro-averaged scores and the highest link-precision. However, the results of BERT are lower by 0.036 macro-F1 score and 0.014 link F1 score. Table 6.9 shows the macro averaged results and the link-specific F1, precision, and recall scores for the nuclear energy and forest dieback datasets.

		F1	P	R
		Nuclear Energy		
Link & No Link	Random	0.496 (0.005)	0.503 (0.000)	0.503 (0.005)
	<code>bilstmAtt</code>	0.356 (0.052)	0.354 (0.004)	0.369 (0.008)
	<code>bilstmAtt_{fine}</code>	0.477 (0.005)	0.483 (0.003)	0.484 (0.004)
	BERT	0.586 (0.008)	0.648 (0.026)	0.615 (0.050)
Link only	Random	0.501 (0.015)	0.504 (0.003)	0.510 (0.014)
	<code>bilstmAtt</code>	0.422 (0.116)	0.390 (0.049)	0.467 (0.028)
	<code>bilstmAtt_{fine}</code>	0.484 (0.027)	0.491 (0.040)	0.492 (0.035)
	BERT	0.485 (0.090)	0.718 (0.150)	0.378 (0.209)
		Forest Dieback		
Link & No Link	Random	0.499 (0.012)	0.507 (0.001)	0.506 (0.004)
	<code>bilstmAtt</code>	0.353 (0.067)	0.348 (0.002)	0.372 (0.002)
	<code>bilstmAtt_{fine}</code>	0.473 (0.005)	0.480 (0.002)	0.483 (0.002)
	BERT	0.550 (0.005)	0.594 (0.005)	0.574 (0.038)
Link only	Random	0.508 (0.018)	0.511 (0.008)	0.517 (0.018)
	<code>bilstmAtt</code>	0.445 (0.137)	0.401 (0.028)	0.505 (0.020)
	<code>bilstmAtt_{fine}</code>	0.508 (0.036)	0.495 (0.028)	0.535 (0.022)
	BERT	0.472 (0.032)	0.638 (0.058)	0.402 (0.178)

Table 6.9: The test results on the nuclear energy and forest dieback datasets.

Nuclear energy dataset When analysing the performance of each model on the nuclear energy dataset in detail, we found that the performance of all models varied greatly between the users. BERT varied between a macro-F1 score of 0.464 (User1) to 0.771 (User4). For User1, User5, and User7, it did not outperform the random baseline. Although the `bilstmAttfine` model only outperformed the random baseline for User0, User1, User4, and User20, it did perform consistently better than the `bilstmAtt`. No model outperformed the random baseline for User5 and User7. Table 6.10 shows the macro-F1 score of all models for each user in the nuclear energy dataset.

Forest dieback dataset On the forest dieback dataset, the results are similar to the nuclear energy dataset. For all users except User1, User3, and User5, BERT reached the highest macro-F1 score. The random baseline is unbeaten for User3 and User5. For User1, the `bilstmAttfine` model performed the best and increased by 0.138 compared with the `bilstmAtt`, and User11 is the only other user in which it outperformed the random baseline. Table 6.10 shows the macro-F1 score of all models for each user in the forest dieback dataset.

	Random	bilstmAtt	bilstmAtt _{fine}	BERT
User0	0.507	0.506	0.509	0.564
User1	0.402	0.289	0.521	0.464
User4	0.496	0.314	0.521	0.771
User5	0.518	0.378	0.461	0.512
User7	0.495	0.360	0.448	0.482
User12	0.502	0.326	0.496	0.674
User13	0.511	0.278	0.387	0.648
User14	0.507	0.339	0.494	0.560
User15	0.522	0.400	0.451	0.632
User16	0.494	0.288	0.463	0.634
User17	0.507	0.394	0.505	0.598
User18	0.478	0.355	0.432	0.552
User19	0.504	0.444	0.482	0.556
User20	0.498	0.319	0.503	0.556

Table 6.10: User-specific results of our models in macro-F1 score on both classes on the nuclear energy dataset.

	Random	bilstmAtt	bilstmAtt _{fine}	BERT
User0	0.504	0.423	0.490	0.584
User1	0.485	0.377	0.515	0.484
User2	0.503	0.326	0.477	0.521
User3	0.507	0.368	0.497	0.471
User4	0.471	0.281	0.440	0.645
User5	0.508	0.318	0.461	0.410
User6	0.508	0.361	0.425	0.627
User7	0.492	0.300	0.455	0.573
User9	0.481	0.422	0.429	0.579
User10	0.502	0.369	0.507	0.598
User11	0.522	0.338	0.512	0.555

Table 6.11: User-specific results of our models in macro-F1 score on both classes on the forest dieback dataset.

6.5 Discussion

Poor performance of the bilstmAtt We found that the bilstmAtt never outperformed the random baseline. To understand the reasons for this, we investigated the training data, and the weakly labelled non-links in particular. We found that for User0, it contained the weakly labelled non-link between the hypothesis “*Nuclear phase not possible due to the profit motive of the companies*” and the piece of evidence “*The information policy is fatal, because it is mostly controlled by a single provider*”. This sample was classified as a link by the bilstmAtt_{fine} model with several randomisation seeds. As a human reader, this link is understandable, because controlling the information policy is a good way to improve profit in times of crisis. A second example is the weakly labelled non-link between the hypothesis “*Rethinking the energy policy is required*” and the piece of evidence “*The security assumptions and security reserves in Japan have not been adequate*”. Again, the statement can be seen as a supporting piece of evidence, thereby as being linked to the hypothesis. This can confuse the model when included in the training data. Therefore, our approach of creating weakly labelled non-links is insufficient. When interacting with real users, strongly labelled non-links are needed to avoid this confusion.

Minor increase due to pre-training Although pre-training a bilstmAtt model on out-of-domain data and then fine-tuning it improved the performance, it still rarely outperformed the random baseline. To understand why, we compared the performance of the bilstmAtt_{fine} model before and after fine-tuning. We found that for most of users, the performance decreased during the fine-tuning process. Therefore, the weakly labelled data also confused the model in the fine-tuning process.

Drop between Development and Test dataset We conducted hyper-parameter tuning by grouping the users based on the amount of data they created and then used representatives as the train/dev dataset. Although our first finding in the bilstmAtt hyper-parameter tuning indicated that the amount of data seems to be barely sufficient, the difference between the development and test datasets is still surprising. Here, even the hyper-parameters seem to be user-specific. When comparing the performance of the bilstmAtt_{fine} model before and after fine-tuning on the development users, the performance increased by about 0.02. Hence, this the effect of the confusing weakly labelled non-links can be reduced by selecting better hyper-parameters. However, the data was too small to learn the user-specific nuances.

Improved performance of BERT compared with the ED task The performance of the out-of-domain trained BERT model was considerably better than all the other models. As in the ED task, we find that BERT is a high-precision but low-recall model. This means, it makes few predictions of links, but the ones it makes are quite accurate. The precision for each user varies between 0.601 (User5) and 0.887 (User4) on the nuclear energy dataset. The recall, on the other hand, is between 0.167 (User1) and 0.516 (User15) on the nuclear energy dataset. On the forest dieback dataset, the precision is between 0.975 (User4) and 0.417 (User5). When looking at the data each user created, the hypotheses of User4 from the nuclear energy dataset created were formulated as a complete claim, for instance “*There are call for alternatives*” or “*They work with a lot of emotions*”. On the same dataset, User5, however, formulated their hypotheses more as categories, such as “*Politics and atomic energy*”, “*Nuclear phase-out especially after Fukushima*”, or “*Criticism*”. Formulating a hypothesis as an actual claim rather than a category is more similar to the controversial statements used as the topic in the out-of-domain training data. However, we could not find a similar distinction between User4 and User5 from the forest dieback dataset.

Upper bound for out-of-domain trained models To determine the upper bound of out-of-domain trained models, we manually labelled the links and non-links of the users for which BERT reached the highest and lowest scores. This means that we selected User4 and User1 from the nuclear energy dataset and User4 and User5 from the forest dieback dataset. Table 6.12 shows that the upper boundaries for an out-of-domain trained EL model in macro-F1 scores are considerably higher than the performance of BERT, leaving room for further improvement. We found no relation between the poor performance of BERT and a low upper bound. However, we found that to distinguish a link from a non-link, we

Nuclear Energy		Forest Dieback	
User1	User4	User4	User5
0.833	0.870	0.868	0.724

Table 6.12: Evaluation of the upper bound of out-of-domain trained EL models for users in the nuclear energy and forest dieback data. The values are macro-F1 scores.

required knowledge about historical events and governments, as well as political positions. For instance, classifying the link between the question “*Can we trust the technology or is it really mature?*” and the statement “*It should be validated that the reactor can handle a power outage.*” correctly requires knowledge about the topic, i.e. nuclear energy, historical events, and what this means in relation to the question. The required historical knowledge is that the Chernobyl disaster started during a simulated power outage and that it means the statement relates to nuclear energy and safety. The connection to the question comes from the knowledge that such a disaster points towards immature technology that should not be trusted.

Fine-tuning BERT on the user data We attempted to further fine-tune BERT on the user-specific data to increase the link-recall. We used the same development data as for the `bilstmAttfine` hyper-parameter optimisation and evaluated different numbers of epochs and learning rates. We found that BERT almost always degenerated to predicting exclusively either the link or the non-link class. We think that the amount of data created by the users is insufficient to further fine-tune a complex model, such as BERT.

6.6 Chapter Conclusion

In this chapter, we investigated different methods of ML to support researchers in history and sociology when it comes to linking pieces of evidence to self-defined hypotheses. To answer our third research question (How well do machine learning-based methods work for EL?), we built different methods for EL. These methods range from directly trained methods, over methods that have been trained on external out-of-domain data to methods that have first been pre-trained on external, out-of-domain data and then fine-tuned on user-specific data. This allowed us to answer our sub-questions as follows:

- ① *Can we train a user-dependent EL model directly from user-created data?*

We found that the amount of data a user creates in a single hour might not be enough to train an EL model with weakly labelled non-links. No directly trained model outperformed a random baseline. However, we found that in creating weakly labelled non-links between hypotheses and pieces of evidence, we sometimes created links that could very well be a positive link. This means that the creation of strongly labelled non-links between hypotheses and pieces of evidence is necessary. Strongly labelled non-links can be created by users rejecting the suggested links between pieces of evidence and hypotheses or by removing an existing link. It is also possible that the amount of data might be too small to build a user-specific EL model.

- ② *Can we use user-independent data for pre-training to reduce the necessary amount of user-specific training data?*

We found that a pre-training an EL model on external out-of-domain data did improve the performance of the EL model. However, the performance during fine-tuning generally deteriorated and dropped from being above the random baseline to being below. We think that this is because of the confusion caused by weakly labelled non-links, which are semantically indistinguishable from links.

- ③ *Does a user-specific EL model outperform a generic one that has been trained on out-of-domain data?*

We found that when training BERT on out-of-domain data, it outperformed both the directly and fine-tuned models. Although it produced the highest macro-F1 scores,

it showed to be a high-precision and low-recall method, leaving room for further improvement in predicting all links.

④ *What is the relation between hyper-parameters and individual users?*

We found that using the same hyper-parameters for users with similar amounts of data resulted in large differences in performance between them. This means that the best-performing hyper-parameters are user-specific and need to be adapted to best support a specific user.

In this chapter, we found that EL is an even more challenging task than ED. We also found that creating random links between hypotheses and previously unrelated pieces of evidence is insufficient and that strongly labelled non-links are necessary, i.e. user-generated non-links between pieces of evidence and hypotheses. Although an out-of-domain trained BERT outperformed the ones trained on strongly labelled links and weakly labelled non-links, we also found that BERT is a high-precision but low-recall method; it does not find many links, but if it predicts one, it is often correct. This means that it would be more applicable to a large corpus rather than a small one that a user can check manually. However, our attempts at further fine-tuning the model failed because of the small amount of available data.

From a long-term perspective, a user-specific model that has been pre-trained on external data and will then be fine-tuned on in-domain strongly labelled links and non-links is the best approach in supporting users. An externally trained model would not be able to adapt to the individual researcher's preferences. However, our finding that the hyper-parameters are user-specific means that the hyper-parameters have to be determined individually for each user. This means that any application aiming at supporting researchers in this task would need to conduct a user-specific hyper-parameter optimisation or additional research in automatically determining the hyper-parameters.

Chapter 7

Interactive Evidence Detection

We established previously that when supporting researchers in the humanities and social sciences, an Interactive Evidence Detection (IED) method must adapt to the individual user. We also demonstrated that only a few documents are required until the model outperforms a well generalising, state-of-the-art method that was trained on out-of-domain data. However, metrics, such as precision, recall, and F1 score, are not always good predictors of whether humans find something useful or not; the practical benefit of such a method might be different from what the intrinsic evaluation indicates. Therefore, to evaluate our fourth research question (RQ 4: How do researchers benefit from interactively trained ED and EL models in their research?), we need to conduct extrinsic evaluations and address the following sub-questions:

- ① Do users perceive the change in quality of ED and EL models while they interact with them?
- ② Do users perceive the suggestions of evidence and which piece of evidence might be linked to a particular hypothesis as beneficial in their work?
- ③ How would researchers use a tool which makes these suggestions in their work?

To answer the sub-questions, we needed to extend EDoHa such that it can train ED and EL models on the data a user creates. To address the cold-start problem, we enabled it to load pre-trained models similar to our transfer-learning approaches in chapter 5 and chapter 6. We also used the ED-ACL-2018-based datasets created in these chapters. We performed the same separation into a detection and linking subtask for the UKP Sentential AM dataset and pre-trained models for them. Furthermore, we machine translated both datasets from English to German to benefit users who work with German text as well.

To evaluate the first three sub-questions, we conducted two user studies:. One employed a group of undergraduate students and the other one employed a group of expert users, i.e. doctoral students, conducting actual research. Each undergraduate student was given the same hypotheses to evaluate and the study took place in a time span of about one hour. Each expert user worked on their respective research subject for about two hours.

7.1 Interactive Evidence Detection with EDoHa

Enabling EDoHa to interactively learn from a user to detect and link evidence required extensions to the user interface, as well as the back-end. Designing these extensions also forced us to consider how a user would interact with EDoHa and how the new features might influence the user’s research.

7.1.1 User Interface

We designed the user interface of EDoHa with two additional goals in mind, namely *unobtrusive suggestions* and *keeping the user in control*.

Unobtrusive suggestions We based our design of the suggestions based on two findings; first, that the behaviour of users varies greatly (chapter 4 and second, that correcting the suggestions of many models may take more effort than labelling by hand (chapter 5). We therefore decided that the presentation of EDoHa’s suggestions should interfere as little as possible with the user’s work. This means that we chose lighter colours for the suggestions than for the user-labelled evidence and that a user is not required to respond immediately or at all to the suggestions; if desired, users can ignore all suggestions and work as they wish.

Keeping the user in control In discussions with historians and social scientists, we found that they have general reservations against losing control over their research methodology. We therefore decided to keep most of the control in the user’s hands. First, by adding a user adjustable confidence threshold to filter out low quality suggestions. Second, by not starting the training automatically, but only upon the user’s request.

Figure 7.1 shows the modified user interface of EDoHa’s document view and figure 7.2 shows the modified evidence linking view. The following items refer to the additions in both the document and evidence linking view:

- ① Suggested pieces of evidence or links between hypotheses and evidence. In the document view, the suggested sentences are highlighted with a light green background colour. If a user wishes to accept a suggestion, they only need to click on it. In the evidence linking view, the suggested pieces of evidence for a hypothesis are added to the user-linked pieces of evidence. The suggestions are also distinct in that they have a smaller font and are also presented in light green colour. If a user wishes to accept a link between a piece of evidence and a hypothesis, they can click the check mark (☑) next to the suggestion; if the user wishes to remove the suggestion they can click the cross (✖).
- ② A confidence threshold slider to reduce the number of suggestions. If the user wishes to see fewer suggestions, they can move the confidence slider to the right which increases the threshold. If the user wishes to see more suggestions, they can move it to the left which decreases the threshold. When moved completely to the left, the threshold will be 50% which is right at the decision boundary of the model. If the user does not wish to see any suggestion, they can set the threshold to 100% which practically turns the suggestions off, because almost no model makes predictions with 100% confidence.
- ③ Users can click the **Train Model** button to trigger the training of either model on all of the user-created data. EDoHa shows a progress bar next to the button to indicate

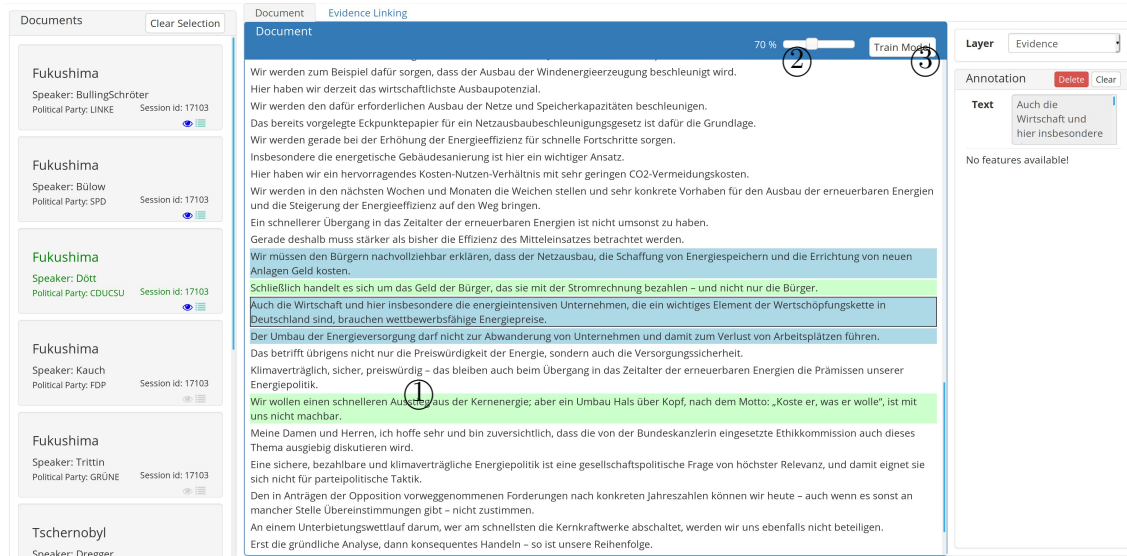


Figure 7.1: Screenshot of EDOHa's document view with the ability to train an ED model. The numbers refer to the additions made to incorporate suggestions.

how the training is proceeding. If the user wishes to see suggestions from the newly trained model, they need to refresh the page.

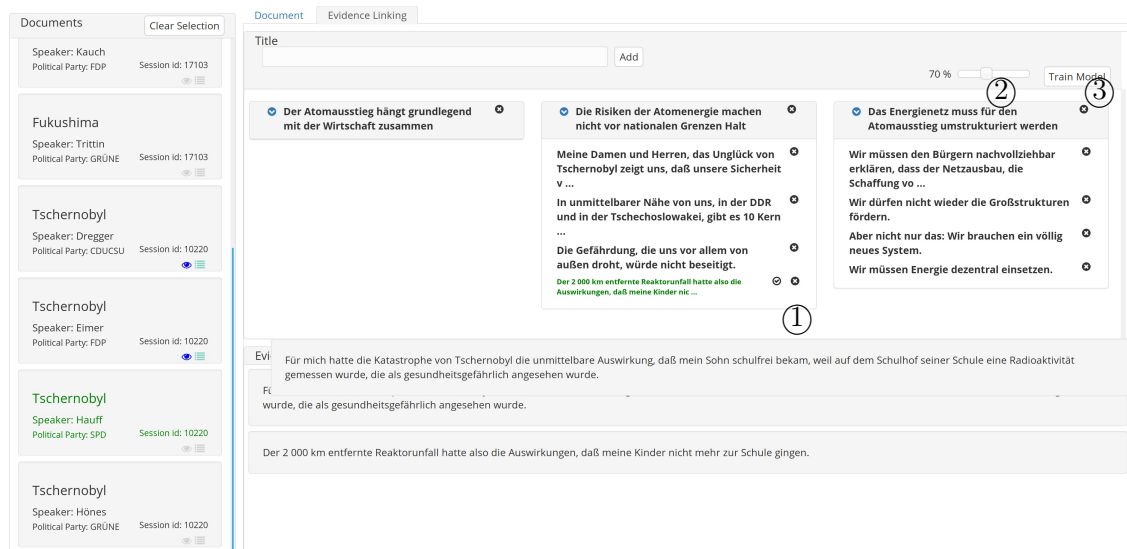


Figure 7.2: Screenshot of the evidence linking view of EDOHa with the ability to train an EL model. The numbers refer to the additions made to incorporate suggestions.

7.1.2 Interactive Training Data Generation

When training an ED model, we first need to construct the training data. Based on our finding that many users work on one document at-a-time, we can create the necessary training data as figure 7.3 illustrates. The user opens the first document and labels sentences that are evidential as evidence. After finishing the first document, the user opens the second one. EDOHa then treats all sentences from the first document and their labels

as training data. After reading and annotating the second document, the user opens the third document and the sentences and labels of the second document are added to the training data. The user can then click on the **Train Model** button which causes EDoHa to use the existing training data to train an ED model. When the user opens the third document, the previously trained ED model is used to make predictions on the sentences of this third document. The user then annotates the third document by accepting suggestions and labelling evidential sentences that have not been suggested. Given that evidence is few and far between, we have to consider that there will be a large class imbalance between evidential and non-evidential sentences. We can address the class imbalance by randomly down-sampling the non-evidential sentences to create balanced training data.

However, not every user may wish to train an ED model only after finishing reading a complete document. This is especially important when the documents are longer than a few pages. In such a case, a user might wish to train an ED model based on the text they have read so far. This means, when training an ED model there are sentences in the documents that the user has not read. We therefore ignore all unlabelled sentences in a document that appear after the last labelled one.

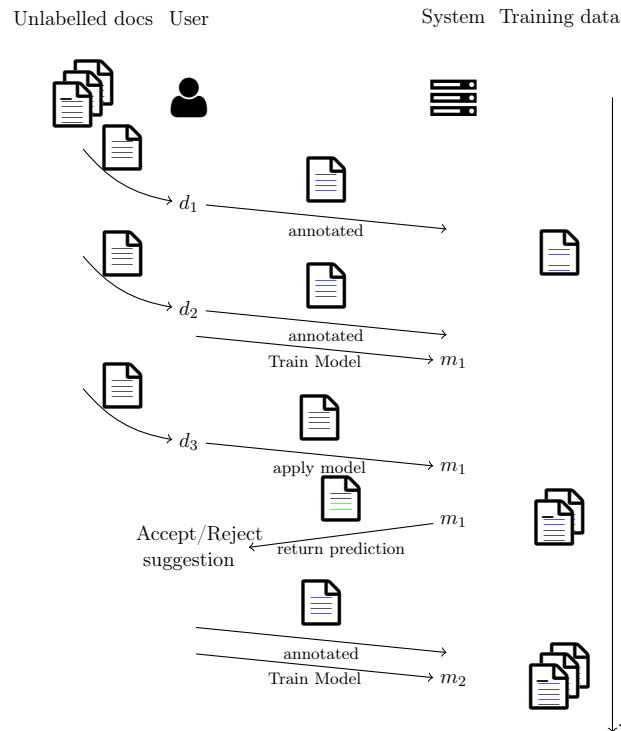


Figure 7.3: After reading the first two documents d_1 and d_2 , the user clicks on **Train Model** to train the first model m_1 . This is then used to suggest evidence on the third document d_3 , which the user corrects and after clicking again on **Train Model**, the system trains the second model m_2 on the training documents d_1 , d_2 , and d_3 .

When training an EL model, we are faced with the problem that users only create positive training data, i.e. links of hypotheses and related pieces of evidence. They do not create any negative training data that contrasts these links from unrelated pieces of evidence and hypotheses, or *non-links*. We address this problem by first creating weakly labelled non-links of pieces of evidence and hypotheses to create a balanced dataset similarly to the creation of a balanced EL dataset in chapter 6. After training the first EL model, the user has the option to accept or reject the suggestions. In the second step, when a user accepts a suggestion, they create a strongly labelled link and if a user rejects one,

they create a strongly labelled non-link between a piece of evidence and a hypothesis. This then reduces the amount of weakly labelled data required to create a balanced dataset and avoids the problem of having irritating training data (see chapter 6). EDoHa also treats all strongly labelled links a user deletes the same. The suggested links are cleared from all strongly labelled non-links. This means, that if a suggested link between a piece of evidence and hypothesis has previously been rejected or deleted by a user, it will be filtered out and not presented in the user interface. The easy creation of strongly labelled non-links can lead to another class imbalance as there are many more possible non-links than links that make sense. We address this problem by letting EDoHa randomly up-sample the links, because we intend to use all strongly labelled data available for training. Figure 7.4 illustrates the creation of strongly and weakly labelled data for the EL task.

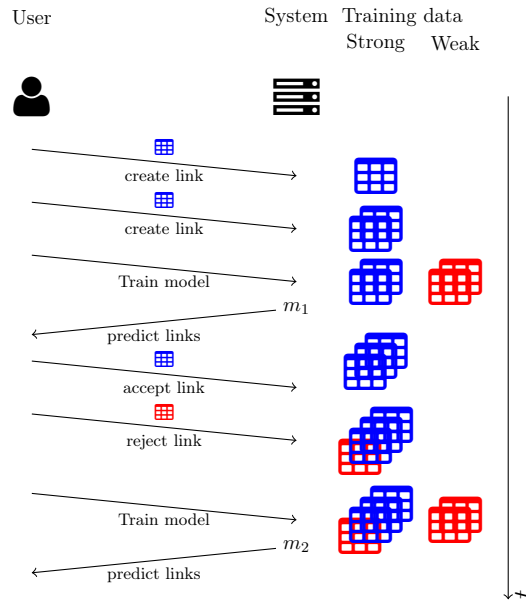


Figure 7.4: The user first creates two links and clicks the **Train Model** button. Then EDoHa generates two weakly labelled non-links and trains the model m_1 which EDoHa then uses to predict links between evidence and titles. After the user accepts one and rejects one link, they again train a new model. To keep the dataset balanced, EDoHa only needs to generate two weakly labelled non-links and train the model m_2 . The blue data represents the links and the red data the non-links.

7.2 Pre-training Data and Models

To address the cold-start problem, we enabled EDoHa to load pre-trained models for both ED as well as EL. We also converted two datasets, one ED dataset and one AM dataset into an appropriate format and machine translated them from English to German.

7.2.1 Datasets

As ED dataset we chose the ED-ACL-2018 dataset and for AM, we chose the UKP Sentential AM dataset. A manual inspection of the translation of the ED-ACL-2018 dataset showed that the translation resulted in semantically identical, grammatically correct sentences. We translated both datasets from English to German through an automatic ma-

chine translation engine.¹ Table 7.1 shows the statistics of the converted ED-ACL-2018 and UKP Sentential AM datasets. The values apply to the English as well as the German versions.

Evidence Detection				
	Split	Topics	Sentences	Evidence
ED-ACL-2018	Train	83	4065	1499
	Test	35	1718	683

Evidence Linking				
	Split	Topics	Links	Non-links
ED-ACL-2018	Train	83	1499	1499
	Test	35	683	683

Table 7.1: Statistics on the datasets used to pre-train the ED and EL models.

Argument Detection				
	Split	Topics	Sentences	Argument
UKP Sentential AM	Train	8	18246	7969
	Dev	8	2059	884
	Test	8	5186	2286

Argument Linking				
	Split	Topics	Links	Non-links
UKP Sentential AM	Train	8	7991	7991
	Dev	8	892	892
	Test	8	2256	2256

Table 7.2: Statistics on the datasets used to pre-train the argument detection and argument linking models.

7.2.2 Model Architectures

For the ED model, we chose the best performing ED model architecture from chapter 5. This means, the ED model is a topic-agnostic bilstm with 100 nodes and a dense layer for classification. This model has no knowledge about any topic or hypothesis the user created and has as input only the candidate sentence.

Based on the good performance of the Enhanced LSTM (ESIM) reported by Hanselowski et al. (2018) we selected it as the ED model architecture. However we found in preliminary experiments on the ED-ACL-2018 dataset that the attention mechanism developed by Q. Chen et al. (2017) did not yield better results than a Siamese bilstm without attention. Instead, we found that a simpler attention mechanism provided a considerable performance benefit.

Our model calculates the attention between the candidate sentence s and topic τ , and

¹Specifically, we used Google Translate <https://translate.google.com/> for the automatic translation.

vice versa, through the sequence of tokens i . First, it builds the contextual token encodings

$$\bar{s}_i = \text{bilstm}(s, i), \quad \forall i \in [1, \dots, \text{len}(s)] \quad (7.1)$$

$$\bar{\tau}_j = \text{bilstm}(\tau, j), \quad \forall j \in [1, \dots, \text{len}(\tau)] \quad (7.2)$$

It then calculates the token-wise attention matrix between the candidate sentence and topic

$$e_{ij} = \bar{s}_i^T \bar{\tau}_j. \quad (7.3)$$

The attention for each token in the candidate sentence (\tilde{s}) is then the sum of the softmax-normalised attention to each token in the topic, with the attention of the topic $\tilde{\tau}$ being calculated accordingly.² This means,

$$\tilde{s}_i = \sum_{j=1}^{\text{len}(\tau)} \frac{\exp(e_{ij})}{\sum_{k=1}^{\text{len}(\tau)} \exp(e_{ik})}, \quad \forall i \in [1, \dots, \text{len}(s)] \quad (7.4)$$

$$\tilde{\tau}_j = \sum_{i=1}^{\text{len}(s)} \frac{\exp(e_{ij})}{\sum_{k=1}^{\text{len}(s)} \exp(e_{ik})}, \quad \forall j \in [1, \dots, \text{len}(\tau)]. \quad (7.5)$$

The model then calculates the attention-weighted contextual token encodings for the candidate sentence $\bar{s}\tilde{s}$ and topic $\bar{\tau}\tilde{\tau}$. For pooling, it builds the sum of the attention-weighted token encodings and selects the encoding of the last token to represent the input sequence. It then concatenates the pooled features. Although Q. Chen et al. (2017) also used other features, such as the attention itself or the difference of the attention and the contextual encodings, we found that neither of them were beneficial when evaluated on the test data of the ED-ACL-2018 dataset. As the final step, our model uses a dense layer for classification. We based our implementation on the ESIM published by Hanselowski et al. (2018). However, we trained the ESIM directly as a classification model with a softmax cross-entropy and an Adam optimiser (Kingma and Ba, 2015), rather than as a ranking model with a Hinge loss, because we treated EL as a classification task and not a ranking task. Our model receives the candidate sentence and a hypothesis as input, as well as their lengths to calculate the attention between them. The attention mechanism does not require learning any additional weights. Both input bilstms use 100 nodes and after concatenating them, we use a dense layer for classification. Figure 7.5 shows the architecture of the EL model.

7.2.3 Evaluation of Pre-trained Models

We evaluated the quality of the pre-trained models by using the train-test-split of the pre-training datasets. For the models trained on the ED-ACL-2018 dataset, we trained them on the dataset's training data and evaluated them on the dataset's testing data. For the models trained on the UKP Sentential AM dataset, we trained them on the training data of all topics and evaluated them on the testing data of all topics.

ED-ACL-2018

For both, ED and EL, we trained the models for 10 epochs with a learning rate of 0.001 and a batch size of 32. We selected these parameters based on few experiments on the ED-ACL-2018 dataset. Although this means the parameters are likely to be slightly overfitted

²Q. Chen et al. (2017) noted that they found no benefit in using a more complex attention mechanism, e.g. by adding a weight vector to each contextual encoding. We therefore implemented their simpler attention mechanism.

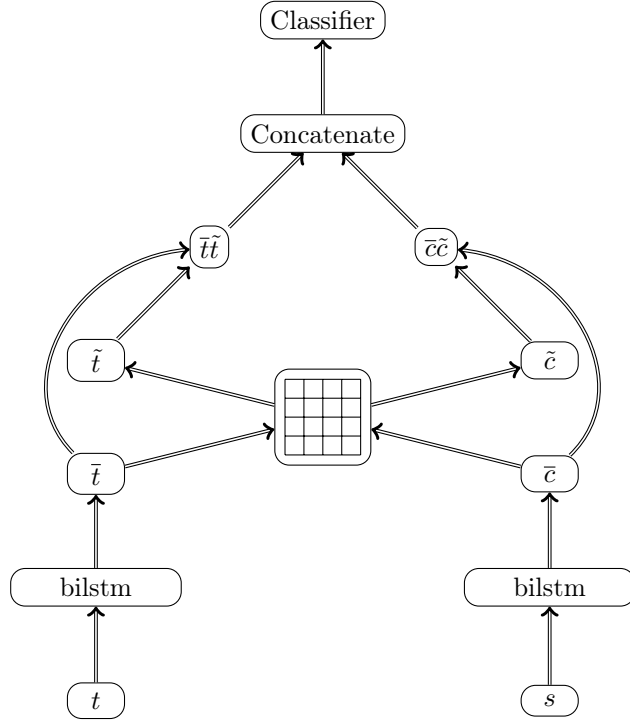


Figure 7.5: Diagram of the ESIM used as EL model in EDoHa.

on the test data, we do not think that this is a large problem, because the models are not meant to be used as is. They are meant to be further fine-tuned by the user, meaning that their initial quality will vary greatly depending on the individual user. All models use 100-dimensional word embeddings as input. For English we selected GloVe embeddings (Pennington, Socher, and Manning, 2014) and for German we selected the embeddings presented by Reimers et al. (2014). Since we are interested in pre-training specific models to be fine-tuned later on, rather than comparing models with one another, we trained them on the single randomisation seed 0. Table 7.3 shows the results of the pre-trained models evaluated on the pre-training testing data.

We found that for ED, the English model reached a macro-F1 score of 0.537 and the German model a macro-F1 score of 0.562. We also found that the translation to German benefited the performance on the evidence class which reached a higher precision, recall, and F1 score. We think there might be two causes for this increase. First, we only evaluated a single randomisation seed and therefore do not know how large the effect of the non-deterministic components is. Second, the machine translation created sentences that are more friendly towards machine learning than human-generated ones.

For the EL task, we found that both macro-F1 score and evidence-F1 score were significantly higher than for the ED model. The English model reached a macro-F1 score of 0.656 and 0.618 evidence-F1 score, whereas the German model reached a macro-F1 score of 0.634 and evidence-F1 score of 0.629. We also found that the translation did affect the macro-precision and macro-recall slightly negatively (-0.034 and -0.26) but benefited the link recall (+0.07).

The results on the English data are considerably lower than the ones presented by Stahlhut (2019a). However, they used 50-dimensional GloVe embeddings, which would have created a mismatch in embeddings size between the English and German models. We therefore decided to use the 100-dimensional embeddings, which were available in English as well as German.

		F1	P	R
Evidence Detection				
EN	Macro	0.537	0.544	0.539
	Evidence	0.399	0.459	0.353
DE	Macro	0.563	0.564	0.567
	Evidence	0.504	0.466	0.549
Random	Macro	0.493	0.493	0.494
	Evidence	0.376	0.389	0.363
Evidence Linking				
EN	Macro	0.656	0.668	0.660
	Link	0.618	0.705	0.551
DE	Macro	0.634	0.634	0.634
	Link	0.629	0.638	0.621
Random	Macro	0.507	0.507	0.507
	Link	0.501	0.507	0.496

Table 7.3: Performance of the pre-trained models for evidence detection and evidence linking in English as well as German.

UKP Sentential AM

For argument detection, we trained the models for 2 epochs, with a learning rate of 0.001 and a batch size of 32. We used the same number of epochs, learning rate, and batch size for the linking task. As word embeddings, we chose the same as for the previous models. Table 7.4 shows the models’ quality for both tasks in English as well as German.

		F1	P	R
Argument Detection				
EN	Macro	0.713	0.733	0.711
	Argument	0.649	0.751	0.571
DE	Macro	0.717	0.718	0.717
	Argument	0.680	0.690	0.671
Random	Macro	0.504	0.504	0.504
	Argument	0.443	0.446	0.440
Argument Linking				
EN	Macro	0.950	0.950	0.950
	Link	0.950	0.946	0.955
DE	Macro	0.932	0.932	0.932
	Link	0.932	0.929	0.935
Random	Macro	0.500	0.500	0.500
	Link	0.502	0.500	0.505

Table 7.4: Performance of the pre-trained models for detection and linking on the AM data.

Both the English and the German argument detection models reached a macro-F1 score of over 0.7. One difference from the ED models is that the translation to German did not affect the model’s performance in macro-F1 negatively; in fact, the German model reached

a slightly higher argument-F1 score than the English one due to a higher argument recall. As with the results of the ED data, we think this increase is due to the larger impact of random noise or an effect of the machine translation process. For argument linking, both the English and the German model reached very high scores with macro-F1 score above 0.9. The translation resulted in a drop in performance of about 0.018 macro-F1 score.

We think that the large difference between the performance of the ED-ACL-2018 and UKP Sentential AM datasets has two reasons. For the detection model, the size of the UKP Sentential AM training data is considerably larger than that of the ED-ACL-2018 dataset (≈ 18.000 compared to ≈ 4000). For linking models, we think that a large factor is the complexity of the topic. In the UKP Sentential AM dataset, the topic consists of one or two tokens, whereas in the ED-ACL-2018 dataset the topic is an entire claim.

7.3 Evaluation

To answer our sub-questions, we conducted two qualitative evaluations, one with a group of undergraduate students participating in a history class and one with three expert users. In the first evaluation we provided the participants with a collection of documents and three hypotheses they were supposed to validate. In the second evaluation we selected a more realistic setting in that the participants provided their own textual sources and hypotheses.

7.3.1 Student Evaluation

We integrated the evaluation into a course on the French Revolution. We scheduled the evaluation at the last lecture of the course and conducted it in our computer lab. Overall, four people participated in the student evaluation.

Evaluation setup We set our evaluation up in a similar fashion than our earlier user studies. This means, we created accounts for each participant beforehand to keep the users anonymous. We also asked them to sign a consent form so that we can further analyse the data they created; the consent form was identical to the ones we used in chapter 4. We gave a short introduction to EDoHa and handed out the description of their task. The users were then asked to log into EDoHa and start evaluating three hypotheses we provided. While the users validated the hypotheses, we answered all emerging questions and supported the users in case of problems with EDoHa. After about 45 minutes, we asked the users to stop working and fill out a questionnaire attached to the task description.

Data As textual sources, we selected the same political speeches on the topic of nuclear energy we used in our first user study. The textual sources were therefore nine speeches from the German parliament, four after the Chernobyl disaster and five after the Fukushima catastrophe.

Student task In contrast to the student tasks in our previous studies, we provided the users with three hypotheses they were supposed to validate, to keep the results more consistent. We selected hypotheses that had been formulated by participants of our first user study, namely “*Nuclear energy is fundamentally connected to the environment*”, “*The risks of nuclear energy do not stop at national borders*”, and “*The power grid needs to be restructured to accommodate the nuclear phase-out*”.³

³All hypotheses were translated from German to English by us.

Questionnaire We revised the questionnaire we developed for our previous user studies to reflect the change in focus now that EDoHa includes IED components. In addition to the previously existing question about whether EDoHa was beneficial to the user and whether they changed their hypotheses, we also included the following binary questions:

- Did you feel that the suggestions were beneficial?
- Did the suggestions influence your decisions?
- Did the tool adapt to your expectation?
- Would you like to continue using EDoHa in your future studies?

Pre-trained models We provided EDoHa with pre-trained models so that it can make suggestions from the beginning. Since the textual sources were in German, we used the German version of the ED-ACL-2018 dataset to pre-train an ED and an EL model. In contrast to the models described above, we used 100-dimensional GloVe embeddings.⁴

7.3.2 Expert User Evaluation

Evaluation setup Three expert users⁵ with a background in sociology or geography participated in a two-hour-long workshop during which they used EDoHa to work on self-selected text. We started with an interactive introduction to get the participants used to EDoHa without any suggestions. We then directed them to set up the pre-trained models by downloading them from a prepared server and uploading them to their project and setting them up. Afterwards, they were able to interactively continue training the ED and EL models. We answered all questions that came up during the workshop and gave additional support, for instance on how to change the project settings to speed up the prediction process in EDoHa. At the end of the workshop, we handed out a questionnaire and collected additional feedback as well as suggestions for improvements.

Data The participants used textual sources from different genres. User1 used three transcripts of interviews they conducted. User2 used four research papers. User2 used two transcripts and a research proposal.

Questionnaire Since the users are researchers, rather than undergraduate students, we revised the questionnaire to reflect this. Hence, the questions were:

1. Did you gain new insights during the exercise?
2. Did you find the tool beneficial for your work?
3. Did you get the impression that the tool’s suggestions adapted to your annotations?
4. Did you find the suggestions for relevant sentences beneficial?
5. Did you find the suggestions for links between evidence and title beneficial?
6. Do you think the tool would be beneficial in speeding up your text annotation?
7. Can you imagine using the system for future research in you PhD project?

⁴This is due to a mistake in pre-training were we accidentally used GloVe embeddings, rather than the German ones.

⁵All users were doctoral candidates from the research training group KRITIS.

The first and last question were binary questions, whereas the other questions used a Likert scale with four levels between yes (1) and no (4). To these quantitative questions, we also added three free-form questions, namely:

8. What did you like about the annotation tool?
9. What did you **dislike** about annotation tool?
10. What kinds of changes would you like to see in EDoHa?

Extension of EDoHa Because two experts used transcripts of interviews, we anticipated a higher rate of words that are not in the vocabulary. Interviews are generally transcribed by hand and this means that spelling mistakes can occur at a higher rate than for intentionally proofread documents. We therefore decided to add an optional rudimentary spelling correction that calculates all permutations of a word with an edit distance of one if the word does not occur in the vocabulary of the embeddings. If a permutation is in the vocabulary, we select it; if no permutation is in the vocabulary, we select the out-of-vocabulary token of the embeddings.

Pre-trained models Since all texts provided by the experts were in English, we only used the English language versions of the pre-trained ED-ACL-2018 datasets.

7.3.3 Results

Student Evaluation

Based on the answers in the questionnaires, we found that three of the four participants felt that the suggestions of EDoHa helped them in validating the hypotheses. All except one user, namely User2, could imagine using EDoHa in their future studies. Table 7.5 shows the responses of all users.

	Beneficial	Influence	User adaptation	Continue use
User1	yes	occasionally	partially	no answer
User2	60%	40%	60%	no answer
User3	no		no answer	yes
User4	yes	no	no answer	yes

Table 7.5: Answers of the student users participating the first user study.

Expert Evaluation

We separated the quantitative questions (questions 1–7) from the qualitative ones to get a deeper understanding of the user’s perceptions and opinions.

Quantitative Responses We found that all experts were interested in continuing to use EDoHa in their research (question 7). They also responded that they thought EDoHa to be beneficial for their work (question 2) and that either the suggestion of evidence (question 3) or the suggestions of links (question 4) were beneficial. User1 found all suggestions to be beneficial, whereas User2 only found the ED model and User3 the EL model to be beneficial. Table 7.6 shows the responses of the users to our binary and Likert scale questions. To keep the responses consistent, we used the values 1 and 4 as values for yes and no of the binary questions.

	Q. 1	Q. 2	Q. 3	Q. 4	Q. 5	Q. 6	Q. 7
User1	1	1	1	1	1	1	1
User2	1	2	3	3	2	2	1
User3	1	2	2	1	4	1	1

Table 7.6: Answers of the expert users to the quantitative questions of the questionnaire. The value 1 corresponds to the answer yes and the value 4 to the answer no.

What the users liked User1 liked the user interface, in particular the Drag & Drop and the confidence slider. User2 liked that EDoHa can be trained to “suit [ones] own concepts” and that it made suggestions that the user may not have considered otherwise. User3 liked the way EDoHa made suggestions, as well as the ability to train the models. User3 also liked the ability to read the text in the application and to create an initial set of annotation categories. This means an initial set of labels that could be assigned to statements in the text, and commented that they could see the models improve.

What the users disliked User2 and User3 criticised the need to switch from the document to the evidence linking view to link a newly found piece of evidence to a hypothesis. User2 concluded from the separation into two views that EDoHa would be more suitable for deductive work, rather than inductive approaches such as *grounded theory*. The users User1 and User2 also mentioned that the response time was sometimes slow. User1 also criticised that EDoHa is not connected to other existing annotation tools and that sometimes, sentences were not segmented correctly. User3 disliked the limitation to sentence-level annotations and would have liked to be able to add annotations for individual words or parts of sentences.

Suggestions from the users All users proposed to include non-evidential sentences into the suggestions of the evidence linking view. User1 extended this idea to also include sentences from documents the user has not yet opened. User2 suggested modifying the user interface so that it is more similar to existing annotation tools, such as MAXQDA⁶ or ATLAS.ti⁷. User3 suggested adding pagination or separating long text into sections and adding an export function to continue working with other annotation tools. The user also suggested allowing the option of working in both views at the same time.

7.4 Discussion

7.4.1 Study Results

Although only one expert user reported having benefited from both the ED and EL suggestions, we think that this is due to the short amount of time the users spent with EDoHa. The expert User2 even noted that there were “not enough annotations” to be beneficial.

The expert users concluded that EDoHa may not be beneficial for particular annotation methods, such as *grounded theory*, but they found EDoHa would benefit them in their work. However, the situations in which EDoHa might be beneficial for them varied greatly. One user concluded EDoHa would benefit them in the beginning of an analysis and another suggested first creating data in a standard annotation tool before importing the data into EDoHa. The third user saw potential in using it for other kinds of discourse analysis, e.g. when analysing newspaper reports related to their research.

⁶<https://www.maxqda.de/>

⁷<https://atlasti.com/de/>

Due to the short time the users interacted with EDoHa, we were unable to ascertain the potential effect of long-term use. One possible effect of long-term use could be that the user and the models converge, i.e. that while the model adapts to the user, the user also adapts to the model. Such an effect can have negative as well as positive consequences. On the negative side, the influence of the tool pushes the user in a direction they are not interested in which could negatively affect their research. However, this is one of the reasons we implemented the *confidence slider* so that the user retains control. On the positive side, the convergence can also create synergy effects. For instance, the expert User2 also noted that they like using EDoHa as a “serendipity tool” because it might suggest links that the user has not thought of themselves. This, in turn enables the user to explore directions they would otherwise have not, thereby extending the user’s research perspective.

7.4.2 Future Work

In both the questionnaire and the discussion after the study, the users made multiple suggestions for enhancements to EDoHa.

Different granularities of annotations Limiting the annotatable pieces of evidence to individual and full sentences is one of the first lessons we learned during the first user study (see chapter 4). However, supporting sub-sentence and multi-sentence annotations is a possible step in the future development of EDoHa.

Relations between hypotheses The current evidence linking view does not support creating relationships between different hypotheses. Although some users suggested including hierarchies or relationships between different hypotheses, we decided against implementing such a feature for now, because it would introduce a large amount of additional complexity to both the user as well as the development. However, we see such a feature and the resulting effects on researchers as future research in IED.

Suggesting non-evidential sentences in the EL view The expert users, in particular, requested the option to have suggestions of unlabelled sentences that could be linked to a particular hypothesis. So far, we decided against the inclusion of such an end-to-end model, because having the user switch between the tasks allows us to simplify the linking task, i.e. *link a piece of evidence to a hypothesis* versus *decide whether a sentence is a piece of evidence regarding a particular hypothesis*. We decided to keep the task simple to have fewer erroneous suggestions. The inclusion of an end-to-end model could result in the users spending the vast majority of their time in the evidence linking view. This can have positive, as well as negative consequences. On the positive side, the user would not need to switch between the view that often, which in turn could further speed up the user’s progress. On the negative side, the user might stop reading the complete source documents and thereby miss a lot of information and context. An open question is how to integrate these suggestions. For example, they can be included exactly like the user-accepted pieces of evidence, or in a different form. However, despite these open questions, the inclusion of an end-to-end model seems like a worthwhile direction for future research, which could have a significant impact on research methodology in the humanities and social sciences.

Evidence re-linking Some users found that sometimes a piece of evidence is suggested to be linked to the wrong hypothesis. They found that although a particular piece of evidence is not relevant for the hypothesis it is suggested, though it is relevant for another one. In such a case, the manual step of rejecting the suggestion, then searching for the

evidence and linking it manually to the relevant hypothesis takes a lot of time; rather than just dragging it from one hypothesis to another. We think that such a feature, especially when also enabling it for existing links between evidence and hypotheses, is a valuable feature for future versions of EDoHa.

Separation of ED and EL When designing EDoHa, we decided to manifest the tasks of ED and EL into two views, namely the *document view* and *evidence linking* view. We made this decision to keep the user interface clear and easy to use. However, this separation into different tasks might have reduced the usefulness to some users by favouring a particular work approach. In future developments, it would be beneficial to reduce this separation so that users would have more freedom in following their personally preferred approach.

Integration into existing annotation tools Several users wrote in their questionnaire, or mentioned in the discussion afterwards, that they would like the idea of having the functionality of EDoHa integrated in their preferred annotation tool. Although this might increase the spread and adoption of modern NLP techniques for researchers in the humanities and social sciences, we think that for evaluation purposes of IED such an integration would not be optimal. First, it would reduce the replicability of our study, because the users would use different user interfaces and possibly different levels of annotation, e.g. token versus sentence level. Second, the integration itself poses considerable challenges with a large amount of uncertainty, such as different levels of access for plug-ins and add-ons. Some annotation tools might offer an interface through which suggestions from ML models can be added, whereas others might not. Furthermore, what kind of additions can be made might also depend on the individual annotation tool. Our final reason not to include EDoHa into existing annotation tools is the large amount of work involved. In choosing an existing open source application and extending it, we had to implement EDoHa once. If we had chosen to build plug-ins for other annotation tools, we would have to implement it multiple times, for different platforms and probably in different programming languages. Our choice of extending an existing open source tool allows for easier replication of our study and avoids the amount of uncertainty regarding the features we can implement.

7.5 Chapter Conclusion

To investigate the question whether the methods developed in previous chapters are beneficial to actual users (RQ 4), we conducted two user studies. One with a small group of students and the other with three researchers in the social sciences.

To conduct the evaluations, we first extended EDoHa so that a user can interactively train ED and EL models. Furthermore, to address the cold-start problem, we enabled EDoHa to load pre-trained ED and EL models and created two datasets for pre-training. We also machine translated them from English to German so that users could work on textual sources in either language. We also found that the machine translation had little effect on the quality of the pre-trained models.

Regarding our sub-questions, we found:

- ① *Do users perceive the change in quality of ED and EL models while they interact with them?*

In both user studies, that the users perceived a change in the quality of the suggestions. The users also noticed that as they created more data, the quality of the model’s prediction increased.

- ② *Do users perceive the suggestions of evidence and which piece of evidence might be linked to a particular hypothesis as beneficial in their work?*

The users who participated in our user studies found the suggestions to be beneficial to their work. However, which part of EDoHa they found beneficial depended on the user. Some expert users found benefit in the suggestions in the ED task while others found it in the EL task. Only one expert user found the suggestions on both tasks to be beneficial.

- ③ *How would researchers use a tool which makes these suggestions in their work?*

The kind of work in which EDoHa would benefit them also varied between the users. Some would use EDoHa in the beginning of a new annotation project while others suggested using it for discourse analysis of newspaper articles. Others suggested using EDoHa as a serendipity tool, because it makes suggestions the user has not thought of before.

The suggestions the users made in the studies also point towards the next steps and challenges ahead for the future development in IED and EDoHa. In particular, the inclusion of an end-to-end method for ED and EL in the evidence linking view, the inclusion of sub-sentence and multi-sentence annotations, or relations between hypotheses.

Chapter 8

Conclusion

Evidence is at the core of most research, whether it is NLP, history, or sociology and creating or finding it takes a vast amount of time. Whereas in NLP we conduct experiments to investigate research questions and test hypotheses, historians and sociologists spend enormous amounts of time on reading source materials. Source materials may be historical documents, news articles, or transcribed interviews. The evidence is then used to draw conclusions and support them.

This thesis investigates how modern techniques in NLP and ML can be used to support these researchers in finding evidence in textual sources. We based our work on two use cases, namely a historian researching the political discourse around nuclear energy and a sociologist researching the effect of cruise ship tourism on critical infrastructure. We mapped their work to NLP tasks to investigate where modern NLP and ML techniques benefit the researcher's work. Figure 8.1 illustrates these tasks. We found the tasks in which to best support researchers with novel research are Evidence Detection (ED) and Evidence Linking (EL). To best support researchers, we first needed to answer four research

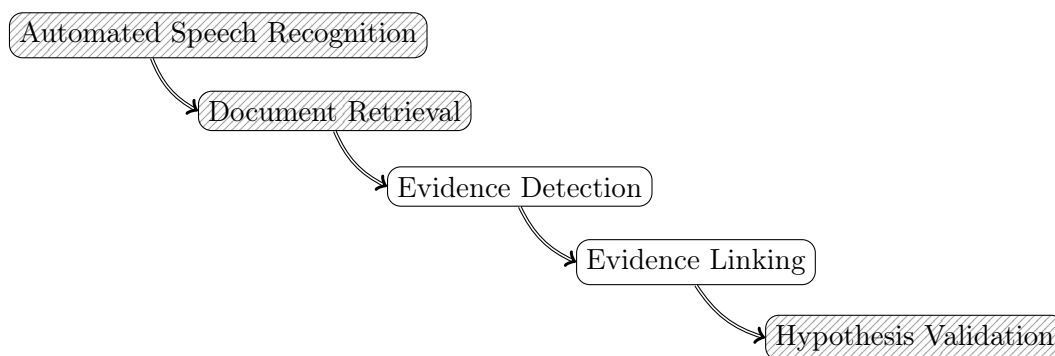


Figure 8.1: The five identified NLP tasks necessary to automatically validate hypotheses in a pipeline model. The tasks with the gray striped background are not considered in this thesis.

questions:

How do researchers in the humanities and social sciences validate their hypotheses?

We investigated our first research question by conducting two user studies with students participating in a teaching-seminar on “*Environmental Catastrophes in the second half*

of the 20th Century". To conduct our studies, we developed an annotation tool named EDoHa which allows a user to label sentences as evidence and link them to self-defined hypotheses. We then performed the user studies with the goal of understanding how they approach their research so that we can build simulations and datasets based on real users' behaviour.

We found that there is no canonical user regarding finding evidence in documents and linking it to hypotheses. Some users work in phases in which they label several sentences as evidence before linking them to one or more hypotheses. Others label a sentence as evidence and immediately link it to one or more hypotheses. We also found that the majority of users worked on the documents from top to bottom as they are presented in EDoHa.

To investigate whether we can build a gold-standard dataset for ED and EL, we calculated the agreement between similar and identical hypotheses. We found no agreement, meaning that supporting researchers in history and sociology requires user-specific models for ED and EL. This is in line with the fact that there are different schools of thought in history as well as sociology meaning that the same text can be read with different eyes. For instance, historiographical materialism interprets the same events differently than subaltern studies.

How well do machine learning-based methods work for ED?

To find the best way to support our users, we needed to understand whether the differences between the users are small enough to be within the margin of a well generalising, state-of-the-art method. We therefore compared the performance of BERT trained on out-of-domain data with smaller models trained interactively on in-domain data. We found that the model interactively trained on in-domain data oftentimes outperforms the out-of-domain trained BERT. Pre-training the interactively trained models on out-of-domain data further improved their performance. We also found that it often takes only one or two documents to do so.

This means that to support researchers in history and sociology, it is best to use the evidence the user finds in the documents to interactively train an ED model. It is even more beneficial, to use out-of-domain data to pre-train these models, so that the user only needs to fine-tune them.

How well do machine learning-based methods work for EL?

Linking pieces of evidence to hypotheses is equally important as finding the evidence itself. But to find the best approach in training the best EL model for a user we are faced with the same questions as in training the best ED model. To investigate these questions we first needed to address the problem that users only create positive links between hypotheses and pieces of evidence. We therefore need to create negative data or non-links if we wish to train a classifier that predicts whether a piece of evidence is linked to a hypothesis or not. We addressed this challenge in creating weakly labelled non-links by randomly pairing a piece of evidence with a hypothesis without duplicating a user-created link.

We then used the strongly labelled links and weakly labelled non-link to compare an out-of-domain trained well generalising, state-of-the-art model with an interactively trained or fine-tuned one. We found that EL is a complicated task in which training directly on the strongly labelled links and weakly labelled non-links rarely performed better than the random baseline. The well generalising but out-of-domain trained model regularly performed better than the random baseline. However, we found that pre-training the EL model and then further fine-tuning it did improve the performance.

We think that this is due to the method in which we created the negative data or non-links. We therefore think that when real users train EL models, it is important to create strongly labelled non-links.

How do researchers benefit from interactively trained ED and EL models in their research?

Based on our findings on the previous research questions, we extended EDoHa to load pre-trained ED and EL models which a user can interactively fine-tune. This means, we integrated the pre-trained and then later further fine-tuned ED model to support users in finding evidence in documents. It also means, that we updated EDoHa to create strongly labelled non-links if they reject a suggestion from the EL model, or if the user deletes a link between a piece of evidence and hypothesis. EDoHa also creates weakly labelled non-links, but only to create a balanced dataset for EL.

We then performed two more user studies with the modified EDoHa, one with a group of students and the other with a group of expert users. We found that expert users in particular found the suggestions to be beneficial. Not only did they notice that the suggestions improved, but they were also interested in continuing to use EDoHa.

However, they also found that EDoHa is more suitable to particular stages of an annotation project or a particular annotation approach. Although all expert users agreed that EDoHa is not very suitable to particular research methods, such as grounded theory, we think that this is due to the separation into ED and EL in the user interface. This distinction facilitates a separation of both tasks and differs from other annotation tools built to support this particular research methodology. However, this is independent from the general benefit which researchers gain from modern NLP and ML methods. They can still benefit in other parts of their work. One user suggested using an external annotation tool to create data to pre-train the ED and EL models. Another user found that the distinction between ED and EL in the user interface makes EDoHa more suitable for deductive approaches in which the hypothesis is defined beforehand and the researcher then starts searching for evidence.

8.1 Lessons Learned

In the research of this thesis we learned several lessons, in particular in ED and ML.

ED cannot be solved by a single model This means that treating ED as a task which can be solved globally might be working for fake news detection and fact-checking, but not in supporting researchers in history and sociology for several reasons. First, there are different schools of thought which influence what kind of hypotheses a researcher formulates and how they interpret the evidence. Second, researchers disagree on which pieces of evidence they select to support similar or identical hypotheses. Third, research in history and especially sociology does not follow a HD approach. It contains considerable inductive components meaning that researchers have not fully formulated all hypotheses before starting their research. For grounded theory in particular, the result is fluid and undefined until the researcher decides to be finished. This means that ED in this context contains a temporal relevance, similar to concept shift and non-stationary Markov Decision Processes.

Metrics aren't everything One considerable finding of this thesis is that although metrics for evaluations are important, they are inadequate in fully informing how well an

ML system works in combination with real users. Both the ED and EL models reached relatively low scores when evaluated against a ground truth. However, even within an hour, our expert users found that the models adapted to the users and that the suggestions were beneficial. Although the users disagreed on the situation in which these suggestions were beneficial, they still found benefit in the suggestions. This means that when considering supporting real users, metrics are only indicators and cannot fully predict the benefit a real user experiences. Therefore, we suggest focussing more on extrinsic evaluations with real human users for ED and EL or any other system which is supposed to support a real user.

Hyper-parameters should be determined automatically In the hyper-parameter optimisation for the EL task, we learned that the parameters are dependent on the individual user. Although this is an interesting finding, it comes with the caveat that each user has to set different hyper-parameters. Of course, we cannot expect users without any background in NLP to conduct their own hyper-parameter optimisation. This means we need to automatically determine the best or at least working hyper-parameters. This also needs to take into account that the hyper-parameters need to be updated as the user creates more data. First, because different amounts of data might need different hyper-parameters; and second, if the user’s goal changes this might also require a different set of hyper-parameters.

8.2 Speeding up Fact-Checking with Interactive Evidence Detection

Although we developed EDoHa with the intention of supporting researchers in the humanities and sociology, there are other fields of work that could benefit from the support that modern NLP offers; most notably, fact-checking.

Similar to conducting research, fact-checking is a very time-consuming task. First, a fact-checker has to determine which claim is worth investigating. Then, they need to find potential sources that support or contradict the claim and afterwards they have to find the actual evidence. Based on this evidence, the fact-checker then decides on the validity of the claim.

First, a fact-checker has to collect relevant documents. Then, they have to read through the documents, gathering evidence related to the claim, until they can come to a verdict. However, while the fact-checker is evaluating a claim, believe in it might already have spread. People who are convinced of the validity of a claim are much harder to convince than somebody who is undecided.

IED offers the possibility of speeding up the fact-checker’s work by reducing the amount of reading they have to do. Interactively learning from the user to find evidence and to link it to hypotheses means the user has to spend less time on reading non-evidential sentences. Suggesting a hypothesis to which a piece of evidence is relevant again reduces the amount of the time the user has to spend creating these links themselves.

Suppose the claim that “*Vaccines contain mercury which causes autism*” is gaining popularity on social media. If this claim is not debunked quickly, many people might believe it’s true and vaccination rates drop; increasing the risk of infection for those who cannot be vaccinated. Debunking this claim could consist of evaluating both components, namely that vaccines contain mercury and that this is harmful. To evaluate the first claim, the fact-checker would search in the databases of medical regulation authorities. Then they find that there is a preservative with the name *Thimerosal* which contains mercury as part

of its molecular structure. It was used to prevent microbial contamination of vaccines in which one flask contains multiple dosages. This means, in cases where multiple vaccination dosages come from the same container, there was a risk of potential contamination of the container. They also find that Thimerosal is no longer used in vaccines for children but may still be used in vaccines for adults.

To evaluate the second claim, that Thimerosal is harmful, the fact-checker then searches through PubMed¹ to find any related published research. The fact-checker then finds multiple medical papers that conducted double blind randomised controlled trials to investigate the health effects of Thimerosal. They then download the abstracts and load them into EDoHa.

In EDoHa the fact-checker then starts reading the abstracts and labels all sentences that contain evidence regarding the health effects of Thimerosal, such as “*The adjusted odds ratios (95% confidence intervals) for ASD [Autism Spectrum Disorder] associated with a 2-SD increase in ethylmercury exposure were 1.12 (0.83–1.51) for prenatal exposure, 0.88 (0.62–1.26) for exposure from birth to 1 month, 0.60 (0.36–0.99) for exposure from birth to 7 months, and 0.60 (0.32–0.97) for exposure from birth to 20 months*”² (①). After reading through one or two abstracts, the fact-checker trains an ED model (②) in EDoHa and then opens the next abstracts. Here, EDoHa already highlights sentences that are similar to the previously labelled ones (③). This allows the fact-checker to find the relevant sentences more quickly than without the suggestions. Figure 8.2 shows EDoHa’s document view with suggestions from previously trained ED models.

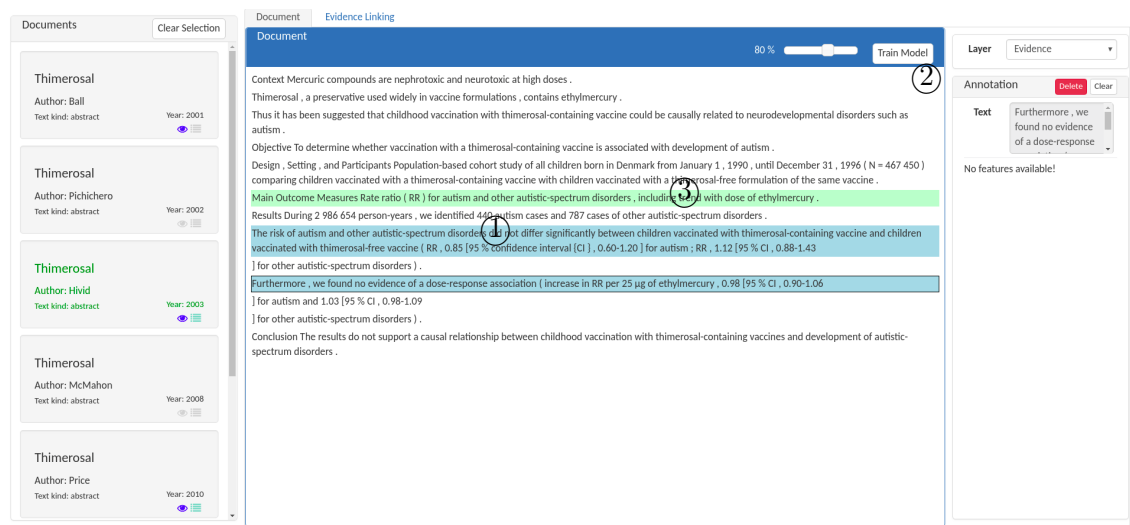


Figure 8.2: The document view of EDoHa with a medical abstract opened and manual annotations, as well as suggestions.

After being satisfied with the collected evidence, the fact-checker changes into the evidence linking view and starts creating groups with the titles: “*Scientific research shows no connection between Thimerosal and autism*” and “*Thimerosal has been shown to be safe when used in vaccines*”. After linking some previously found evidence to these groups (①), the fact-checker then trains an EL model (②). Afterwards, some groups already contain suggestions for relevant pieces of evidence (③) which the fact-checker can then accept or reject. This allows the fact-checker to work faster until they come to an assessment and

¹<https://www.ncbi.nlm.nih.gov/pubmed/>

²This example is taken from McMahon et al. (2008)

write the appropriate report. Figure 8.3 shows how EDoHa’s linking view can be used by the fact-checker.

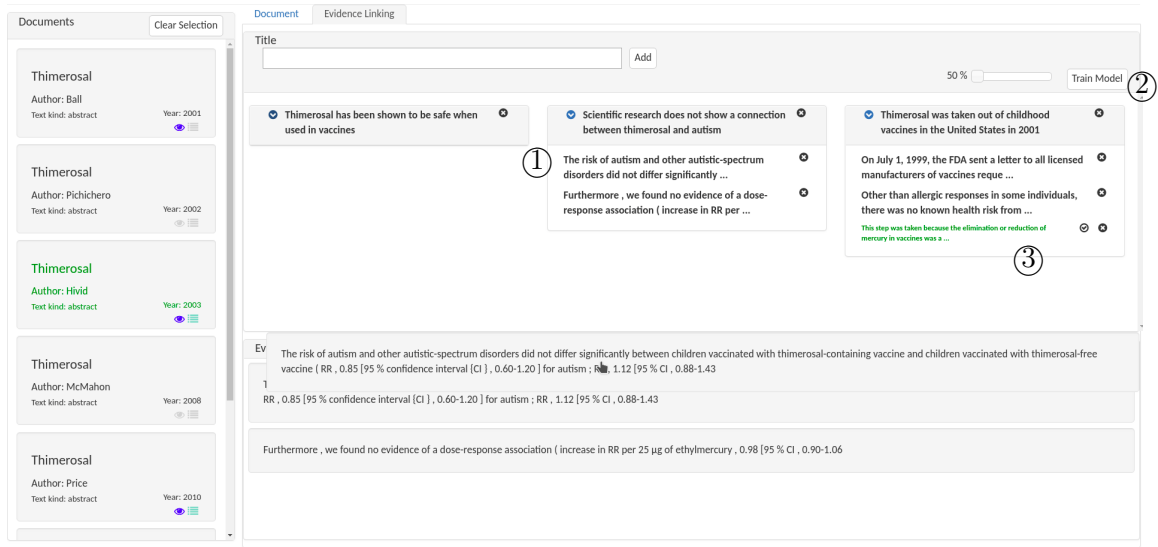


Figure 8.3: The evidence linking view of EDoHa with manual links and suggested links between claims and pieces of evidence. The numbers refer to the individual components described in the main text.

8.3 Future Work

The research conducted in this thesis opens up several directions both for research in NLP and interdisciplinary research. On the NLP side, there are two immediate challenges ahead, namely the improvement of the ED and EL methods and the inclusion of multi-sentence evidence. On the interdisciplinary side, there are challenges ahead in improving the direct benefit of researchers in history and sociology.

Improvement of ED and EL methods The clearest challenge is the improvement of the performance of the ED and EL models. One direction is the inclusion of *small data* ML methods to improve the quality of the interactively trained models. A second direction could be to leverage similarities between users. This could be done either through approaches such as *collaborative filtering* or via *parameter sharing*. In collaborative filtering, if a user labels a piece of evidence which multiple other users labelled as well, the user might be labelling other pieces of evidence shared with the other users. For parameter sharing, it might be possible to treat user-dependent ED or EL as a multi-task learning problem. Each user would be treated as a separate task. Another approach would be to include active learning components in the user interface. If knowing the label of a specific sentence would be particularly beneficial for the model, it could request the user to label this one either as evidence or not. This approach could also be adapted to the EL model.

Automatic selection of hyper-parameters Especially in our experiments in supporting individual researchers in linking their evidence to hypotheses, we found that the hyper-parameters did not depend on the amount of available data, but the nuances of the individual user. This opens up a new direction for hyper-parameter optimisation in which hyper-parameters need to be determined without having a separate development dataset.

This is especially important in supporting researchers in history and sociology, because it seems infeasible to perform an independent hyper-parameter optimisation for each task, let alone each user.

Multi-sentence evidence Several users requested allowing the labelling of multiple sentences as a single piece of evidence. This would change the kind of task from classification to sequence tagging. However, if the particular sentences are following one another in the document, we can treat it as a sequential decision-making task. This means that based on the previously labelled pieces of evidence, we could build a classifier that gives positive feedback if and only if the provided sequence of sentences is a piece of evidence. If we enable delaying the decision for sentences being evidential or not, depending on the next sentence, we can use reinforcement learning to train a sequence tagger with a classifier providing the basis for the reward function. If the sentences that in combination form a piece of evidence must not follow each other directly, other methods for combinatorial optimisation can be applied, e.g. genetic algorithms.

Improvements in the user interface of EDoHa As indicated by the participants of our user studies, the user interface of EDoHa can be improved to better support historians and sociologists. This can be the integration of an end-to-end ED and EL model either as a pipeline or a joint method. It would also improve the support for researchers if it were possible to create connections between hypotheses. Another approach to improving the amount of support EDoHa delivers would be to break up the separation of ED and EL. For instance, when labelling a piece of evidence, it would be possible to already suggest a hypothesis this particular piece of evidence can be linked to. It is also possible to completely remove the distinction between the two tasks.

Integration of large corpora as data sources Another interdisciplinary improvement is the integration of an external document retrieval system into EDoHa. This would allow a user to first query for relevant documents and if a particular document is interesting, it could be imported into EDoHa. Although this is more of a technical challenge, it would still improve its support for researchers in history, especially if the document retrieval system has access to a large corpus of historically important documents.

Support of hypotheses over time Finally, whereas this research focusses on supporting researchers in history in particular, it does not consider the temporal dimension of the available data. For instance, the document creation time containing a particular piece of evidence can be of great importance. Including this meta-datum would allow creating graphs in which the number of pieces of evidence related to a particular hypothesis are shown over time. This would allow finding a point in time in which the discourse changed, similarly as the discourse changed with the second wave of feminism. However, this research is particularly challenging, as it would need to aggregate pieces of evidence which are semantically identical, but syntactically different.

Bibliography

- Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Rutu Rinott, Dan Gutfreund, and Noam Slonim (June 2014). “A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics”. In: *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, pp. 64–68. DOI: 10.3115/v1/W14-2109. URL: <https://www.aclweb.org/anthology/W14-2109>.
- Ajjour, Yamen, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein (2019). “Data Acquisition for Argument Search: The Args.Me Corpus”. In: *KI 2019: Advances in Artificial Intelligence*. Ed. by Christoph Benzmüller and Heiner Stuckenschmidt. Cham: Springer International Publishing, pp. 48–59. ISBN: 978-3-030-30179-8. URL: https://link.springer.com/chapter/10.1007/978-3-030-30179-8_4.
- Alhindi, Tariq, Savvas Petridis, and Smaranda Muresan (Nov. 2018). “Where Is Your Evidence: Improving Fact-Checking by Justification Modeling”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, pp. 85–90. URL: <https://www.aclweb.org/anthology/W18-5513> (visited on 07/18/2019).
- American Sociological Association (2008). *The Field of Sociology*. An Introduction to Sociology. URL: <https://www.asanet.org/sites/default/files/savvy/introtosociology/Documents/Field%20of%20sociology033108.htm> (visited on 03/05/2020).
- Andersen, Hanne and Brian Hepburn (2016). “Scientific Method”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2016. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/sum2016/entries/scientific-method/>.
- Andrews, Pierre, Marco De Boni, Suresh Manandhar, and Marco De (2006). “Persuasive Argumentation in Human Computer Dialogue.” In: *AAAI Spring Symposium: Argumentation for Consumers of Healthcare*, pp. 8–13. URL: <http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-01/SS06-01-002.pdf> (visited on 08/08/2016).
- Artstein, Ron and Massimo Poesio (Dec. 2008). “Inter-Coder Agreement for Computational Linguistics”. In: *Computational Linguistics* 34.4, pp. 555–596. ISSN: 0891-2017. DOI: 10.1162/coli.07-034-R2. URL: <http://dx.doi.org/10.1162/coli.07-034-R2>.
- Atkeson, Christopher G. and Stefan Schaal (1997). “Robot Learning from Demonstration”. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. ICML ’97. Vol. 97. Morgan Kaufmann Publishers Inc., pp. 12–20.

- Baker, Alan (2016). “Simplicity”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/win2016/entries/simplicity/>.
- Baldi, P., S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen (May 1, 2000). “Assessing the Accuracy of Prediction Algorithms for Classification: An Overview”. In: *Bioinformatics* 16.5, pp. 412–424. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/16.5.412. URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/16.5.412> (visited on 11/27/2019).
- Baur, Nina and Hubert Knoblauch (2018). “Die Interpretativität des Quantitativen”. In: *Soziologie* 47. JG.4, pp. 439–461.
- Boltužić, Filip and Jan Šnajder (June 2015). “Identifying Prominent Arguments in On-line Debates Using Semantic Textual Similarity”. In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Denver, CO: Association for Computational Linguistics, pp. 110–115. DOI: 10.3115/v1/W15-0514. URL: <http://www.aclweb.org/anthology/W15-0514>.
- Burdick, Anne, ed. (Nov. 2012). *Digital Humanities*. Cambridge, MA: MIT Press. 141 pp. ISBN: 978-0-262-01847-0.
- Burges, Chris J.C. (June 2010). *From RankNet to LambdaRank to LambdaMART: An Overview*. MSR-TR-2010-82. URL: <https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/>.
- Carnap, Rudolf (1983). “The Concept of Confirming Evidence”. In: *The Concept of Evidence*. Oxford Readings in Philosophy. New York, NY, USA: Oxford University Press, pp. 79–94. ISBN: 0-19-875062-5.
- Chakrabarty, Dipesh (2000). “Subaltern Studies and Postcolonial Historiography”. In: *Nepantla: Views from South* 1.1, pp. 9–32. DOI: 10.4135/9781848608238.n14. URL: http://sk.sagepub.com/reference/hdbk_historicalsoc/n14.xml (visited on 03/04/2020).
- Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen (July 2017). “Enhanced LSTM for Natural Language Inference”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1657–1668. DOI: 10.18653/v1/P17-1152. URL: <http://aclweb.org/anthology/P17-1152>.
- Cohen, Jacob (Apr. 1960). “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1, pp. 37–46. ISSN: 0013-1644, 1552-3888. DOI: 10.1177/001316446002000104. URL: <http://journals.sagepub.com/doi/10.1177/001316446002000104> (visited on 11/26/2019).
- Cohnitz, Daniel and Marcus Rossberg (2019). “Nelson Goodman”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/sum2019/entries/goodman/>.
- Collins, Michael and Nigel Duffy (July 2002). “New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 263–270. DOI: 10.3115/1073083.1073128. URL: <https://www.aclweb.org/anthology/P02-1034>.

- Daston, Lorraine (1991). “Marvelous Facts and Miraculous Evidence in Early Modern Europe”. In: *Critical Inquiry* 18.1, pp. 93–124. DOI: 10.1086/448625.
- Daxenberger, Johannes, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych (Sept. 2017). “What Is the Essence of a Claim? Cross-Domain Claim Identification”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2055–2066. DOI: 10.18653/v1/D17-1218. URL: <https://www.aclweb.org/anthology/D17-1218>.
- Dayan, Peter (1992). “The Convergence of TD(X) for General X”. In: *Machine Learning* 8, pp. 341–363.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (Oct. 10, 2018). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs]. URL: <http://arxiv.org/abs/1810.04805> (visited on 03/22/2019).
- Dorr, Bonnie, Christof Monz, Stacy President, Richard Schwartz, and David Zajic (June 2005). “A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?” In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 1–8. URL: <https://www.aclweb.org/anthology/W05-0901>.
- Eckart de Castilho, Richard and Iryna Gurevych (Aug. 2014). “A Broad-Coverage Collection of Portable NLP Components for Building Shareable Analysis Pipelines”. In: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1–11. DOI: 10.3115/v1/W14-5201. URL: <https://www.aclweb.org/anthology/W14-5201> (visited on 08/07/2019).
- Engels, Jens Ivo, Kristof Lukitsch, Marcel Müller, and Chris Stahlhut (Aug. 2018a). “Criticality”. In: *Key Concepts for Critical Infrastructure Research*. Springer VS, Wiesbaden, pp. 11–20. ISBN: 978-3-658-22919-1 978-3-658-22920-7. URL: <http://tubiblio.ulb.tu-darmstadt.de/106805/>.
- Engels, Jens Ivo, Kristof Lukitsch, Marcel Müller, Chris Stahlhut, Stephanie Eifert, Alice Knauf, Nadja Thiessen, Ivonne Elsner, Andreas Huck, Manas Marathe, Arturo Crespo, Marcus Dombois, and Jan Henning (Aug. 2018b). “Relations between the Concepts”. In: *Key Concepts for Critical Infrastructure Research*. Springer VS, Wiesbaden, pp. 45–52. ISBN: 978-3-658-22919-1 978-3-658-22920-7. URL: <http://tubiblio.ulb.tu-darmstadt.de/106805/>.
- Engels, Jens Ivo, Jochen Monstadt, Marcus Dombois, Sybille Frank, Chris Stahlhut, and Tina Enders (2019). “Urban Infrastructures: Criticality, Vulnerability and Protection. Report of the International Conference of the Research Training Group KRITIS at Technische Universität Darmstadt, Germany”. In: *Urban Infrastructures: Criticality, Vulnerability and Protection*. Ed. by Jens Ivo Engels. Darmstadt, Germany. URL: <http://tubiblio.ulb.tu-darmstadt.de/113656/>.
- Ferreira, William and Andreas Vlachos (June 2016). “Emergent: A Novel Data-Set for Stance Classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

- gies. San Diego, California: Association for Computational Linguistics, pp. 1163–1168. DOI: 10.18653/v1/N16-1138. URL: <https://www.aclweb.org/anthology/N16-1138>.
- Feynman, Richard (1985). *The Character of Physical Law*. 12. printing. The MIT Press Paperback Series 66. Cambridge, Mass.: MIT Press. 173 pp. ISBN: 0 262 56003 8.
- Forman, George and Martin Scholz (Nov. 2010). “Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement”. In: *SIGKDD Explor. Newsl.* 12.1, pp. 49–57. ISSN: 1931-0145. DOI: 10.1145/1882471.1882479. URL: <http://doi.acm.org/10.1145/1882471.1882479> (visited on 01/21/2016).
- Frau, Johanna, Milagro Teruel, Laura Alonso Alemany, and Serena Villata (May 2019). “Different Flavors of Attention Networks for Argument Mining”. In: *FLAIRS 2019 - 32th International Florida Artificial Intelligence Research Society Conference*, Sarasota, United States, p. 6.
- Gao, Yang, Christian M. Meyer, and Iryna Gurevych (Nov. 2018). “APRIL: Interactively Learning to Summarise by Combining Active Preference Learning and Reinforcement Learning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Vol. Long Papers. URL: <http://tubiblio.ulb.tu-darmstadt.de/106848/>.
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (Apr. 2, 2018). “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes”. In: *Proceedings of the National Academy of Sciences*, p. 201720347. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720347115. pmid: 29615513. URL: <http://www.pnas.org/content/early/2018/03/30/1720347115> (visited on 04/06/2018).
- Geuss, Raymond (1981). *The Idea of a Critical Theory: Habermas and the Frankfurt School*. Cambridge University Press.
- Głowacka, Dorota, Tuukka Ruotsalo, Ksenia Konuyshkova, kumaripaba Athukorala, Samuel Kaski, and Giulio Jacucci (2013). “Directing Exploratory Search: Reinforcement Learning from User Interactions with Keywords”. In: *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. IUI ’13. New York, NY, USA: ACM, pp. 117–128. ISBN: 978-1-4503-1965-2. DOI: 10.1145/2449396.2449413. URL: <http://doi.acm.org/10.1145/2449396.2449413> (visited on 05/12/2017).
- Goodman, Nelson (1983). “Prospects for a Theory of Projection”. In: *The Concept of Evidence*. Oxford Readings in Philosophy. New York, NY, USA: Oxford University Press, pp. 63–78. ISBN: 0-19-875062-5.
- Gretz, Shai, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim (Nov. 26, 2019). *A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis*. arXiv: 1911.11408 [cs]. URL: <http://arxiv.org/abs/1911.11408> (visited on 03/11/2020).
- Habernal, Ivan, Judith Eckle-Kohler, and Iryna Gurevych (July 2014). “Argumentation Mining on the Web from Information Seeking Perspective”. In: *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*. Vol. 1341. Forlì-Cesena, Italy: CEUR Workshop Proceedings, p. 14.
- Habernal, Ivan and Iryna Gurevych (Nov. 2016a). “What Makes a Convincing Argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language*

- Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1214–1223. DOI: 10.18653/v1/D16-1129. URL: <https://www.aclweb.org/anthology/D16-1129>.
- (Aug. 2016b). “Which Argument Is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1589–1599. DOI: 10.18653/v1/P16-1150. URL: <https://www.aclweb.org/anthology/P16-1150>.
- Hanselowski, Andreas and Iryna Gurevych (Nov. 2017). “A Framework for Automated Fact-Checking for Real-Time Validation of Emerging Claims on the Web”. In: *NIPS 2017 Workshop on Prioritising Online Content*. Long Beach, USA. URL: https://www.k4all.org/wp-content/uploads/2017/09/WPOC2017_paper_6.pdf.
- Hanselowski, Andreas, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych (Nov. 2019). “A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 493–503. DOI: 10.18653/v1/K19-1046. URL: <https://www.aclweb.org/anthology/K19-1046>.
- Hanselowski, Andreas, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych (Nov. 2018). “Multi-Sentence Textual Entailment for Claim Verification”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, pp. 103–108. URL: <http://www.aclweb.org/anthology/W18-5516>.
- Hanson, Norwood Russell (1983). “The Logic of Discovery”. In: *The Concept of Evidence*. Oxford Readings in Philosophy. New York, NY, USA: Oxford University Press, pp. 53–62. ISBN: 0-19-875062-5.
- Hempel, Carl G. (1983). “Studies in the Logic of Confirmation”. In: *The Concept of Evidence*. Oxford Readings in Philosophy. New York, NY, USA: Oxford University Press, pp. 10–43. ISBN: 0-19-875062-5.
- Howard, Jeremy and Sebastian Ruder (Jan. 18, 2018). *Universal Language Model Fine-Tuning for Text Classification*. arXiv: 1801.06146 [cs, stat]. URL: <http://arxiv.org/abs/1801.06146> (visited on 05/25/2018).
- Hua, Xinyu and Lu Wang (July 2017). “Understanding and Detecting Supporting Arguments of Diverse Types”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 203–208. URL: <http://aclweb.org/anthology/P17-2032>.
- Jordan, Stefan (Feb. 11, 2010). “Vetorecht Der Quellen”. In: *Docupedia-Zeitgeschichte* 1.0. DOI: <http://dx.doi.org/10.14765/zzf.dok.2.570.v1>. URL: http://docupedia.de/zg/jordan_vetorecht_quellen_v1_de_2010.
- Kangasrääsiö, Antti, Yi Chen, Dorota Glowacka, and Samuel Kaski (2016). “Interactive Modeling of Concept Drift and Errors in Relevance Feedback”. In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. UMAP ’16. New York, NY, USA: ACM, pp. 185–193. ISBN: 978-1-4503-4368-8. DOI: 10.1145/2930238.2930243. URL: <http://doi.acm.org/10.1145/2930238.2930243> (visited on 05/10/2017).

- Kasai, Jungo, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa (July 2019). “Low-Resource Deep Entity Resolution with Transfer and Active Learning”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 5851–5861. DOI: 10.18653/v1/P19-1586. URL: <https://www.aclweb.org/anthology/P19-1586> (visited on 09/24/2019).
- Kelly, Thomas (2016). “Evidence”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/win2016/entries/evidence/> (visited on 07/17/2019).
- Al-Khatib, Khalid, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein (Dec. 2016). “A News Editorial Corpus for Mining Argumentation Strategies”. In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING 16)*, pp. 3433–3443. URL: <http://www.aclweb.org/anthology/C/C16/C16-1324.pdf> (visited on 05/17/2017).
- King, Gary and Langche Zeng (2001). “Logistic Regression in Rare Events Data”. In: *Political Analysis* 9.2, pp. 137–163. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/oxfordjournals.pan.a004868. URL: <https://www.cambridge.org/core/journals/political-analysis/article/logistic-regression-in-rare-events-data/1E09F0F36F89DF12A823130FDF0> (visited on 05/13/2019).
- Kingma, Diederik P. and Jimmy Ba (May 2015). “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. San Diego, CA. URL: <http://arxiv.org/abs/1412.6980>.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (Aug. 2018). “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation”. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 5–9. URL: <https://www.aclweb.org/anthology/C18-2002>.
- Landis, J. Richard and Gary G. Koch (1977). “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1, pp. 159–174. ISSN: 0006341X, 15410420. JSTOR: 2529310.
- Lee, Ji-Ung, Christian M. Meyer, and Iryna Gurevych (July 2020). “Empowering Active Learning to Jointly Optimize System and User Demands”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4233–4247. DOI: 10.18653/v1/2020.acl-main.390. URL: <https://www.aclweb.org/anthology/2020.acl-main.390>.
- Liebeck, Matthias, Katharina Esau, and Stefan Conrad (Aug. 2016). “What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld”. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. 3rd Workshop on Argument Mining ACL 2016, Berlin. Berlin, Germany: Association for Computational Linguistics, pp. 144–153. URL: <http://www.aclweb.org/anthology/W16-2817>.

- Liga, Davide (Aug. 2019). “Argumentative Evidences Classification and Argument Scheme Detection Using Tree Kernels”. In: *Proceedings of the 6th Workshop on Argument Mining*. Florence, Italy: Association for Computational Linguistics, pp. 92–97. URL: <https://www.aclweb.org/anthology/W19-4511> (visited on 07/31/2019).
- Lin, Bill Yuchen, Dong-Ho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren (July 2019). “AlpacaTag: An Active Learning-Based Crowd Annotation Framework for Sequence Tagging”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 58–63. DOI: 10.18653/v1/P19-3010. URL: <https://www.aclweb.org/anthology/P19-3010> (visited on 09/24/2019).
- Lin, Chin-Yew (July 2004). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Ed. by Stan Szpakowicz Marie-Francine Moens. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81.
- Lippi, Marco and Paolo Torroni (Dec. 15, 2016). “MARGOT: A Web Server for Argumentation Mining”. In: *Expert Systems with Applications* 65, pp. 292–303. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2016.08.050. URL: <https://www.sciencedirect.com/science/article/pii/S0957417416304493> (visited on 05/28/2019).
- Little, Daniel (2017). “Philosophy of History”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2017. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/sum2017/entries/history/>.
- Lloyd, Christopher (Oct. 24, 2008). “Historiographic Schools”. In: *A Companion to the Philosophy of History and Historiography*. Ed. by Aviezer Tucker. Oxford, UK: Wiley-Blackwell, pp. 371–380. ISBN: 978-1-4051-4908-2. DOI: 10.1002/9781444304916.ch33. URL: <http://doi.wiley.com/10.1002/9781444304916.ch33> (visited on 02/24/2020).
- Lughofer, Edwin (Aug. 2011). “Hybrid Active Learning for Reducing the Annotation Effort of Operators in Classification Systems”. In: *Pattern Recognition* 45.2, pp. 884–896. ISSN: 00313203. DOI: 10.1016/j.patcog.2011.08.009. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0031320311003463> (visited on 03/13/2020).
- Lukin, Stephanie, Pranav Anand, Marilyn Walker, and Steve Whittaker (Apr. 2017). “Argument Strength Is in the Eye of the Beholder: Audience Effects in Persuasion”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 742–753. URL: <https://nlds.soe.ucsc.edu/sites/default/files/belief-eacl-v12.pdf> (visited on 03/06/2017).
- Ma, Jing, Wei Gao, Shafiq Joty, and Kam-Fai Wong (July 2019). “Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2561–2571. DOI: 10.18653/v1/P19-1244. URL: <https://www.aclweb.org/anthology/P19-1244>.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (July 7, 2008). *Introduction to Information Retrieval*. Anniversary. New York: Cambridge University Press. 506 pp. ISBN: 978-0-521-86571-5.
- Matthews, B.W. (1975). “Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme”. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure*

- 405.2, pp. 442–451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9. URL: <http://www.sciencedirect.com/science/article/pii/0005279575901099>.
- Mayer, Tobias, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata (2018). “Argument Mining on Clinical Trials”. In: *Computational Models of Argument: Proceedings of COMMA 2018*. Vol. 305. Frontiers in Artificial Intelligence and Applications. Warsaw, Poland: IOS Press, pp. 137–148. DOI: 10.3233/978-1-61499-906-5-137.
- Mayer, Tobias, Elena Cabrio, and Serena Villata (Nov. 2018). “Evidence Type Classification in Randomized Controlled Trials”. In: *Proceedings of the 5th Workshop on Argument Mining*. Brussels, Belgium: Association for Computational Linguistics, pp. 29–34. URL: <https://www.aclweb.org/anthology/W18-5204>.
- McMahon, A. W., J. K. Iskander, P. Haber, M. M. Braun, and R. Ball (Jan. 17, 2008). “Inactivated Influenza Vaccine (IIV) in Children <2 Years of Age: Examination of Selected Adverse Events Reported to the Vaccine Adverse Event Reporting System (VAERS) after Thimerosal-Free or Thimerosal-Containing Vaccine”. In: *Vaccine* 26.3, pp. 427–429. ISSN: 0264-410X. DOI: 10.1016/j.vaccine.2007.10.071. pmid: 18093701.
- Misra, Amita, Brian Ecker, and Marilyn Walker (Sept. 2016). “Measuring the Similarity of Sentential Arguments in Dialogue”. In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, pp. 276–287. DOI: 10.18653/v1/W16-3636. URL: <https://www.aclweb.org/anthology/W16-3636>.
- Mollá, Diego and Ben Hutchinson (Apr. 2003). “Intrinsic versus Extrinsic Evaluations of Parsing Systems”. In: *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?* Columbus, Ohio: Association for Computational Linguistics, pp. 43–50. URL: <https://www.aclweb.org/anthology/W03-2806>.
- Orbach, Matan, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim (Nov. 2019). “A Dataset of General-Purpose Rebuttal”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5595–5605. DOI: 10.18653/v1/D19-1561. URL: <https://www.aclweb.org/anthology/D19-1561>.
- P.V.S., Avinash and Christian M. Meyer (July 2017). “Joint Optimization of User-Desired Content in Multi-Document Summaries by Learning from User Feedback”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1353–1363. URL: <http://aclweb.org/anthology/P17-1124>.
- Peldszus, Andreas and Manfred Stede (2013). “From Argument Diagrams to Argumentation Mining in Texts: A Survey”. In: *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7.1, pp. 1–31. URL: <http://www.igi-global.com/article/from-argument-diagrams-to-argumentation-mining-in-texts/87173> (visited on 03/24/2017).
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Process-*

- ing (EMNLP), pp. 1532–1543. URL: <http://www.aclweb.org/anthology/D14-1162> (visited on 12/12/2016).
- Plesser, Hans E. (Jan. 18, 2018). “Reproducibility vs. Replicability: A Brief History of a Confused Terminology”. In: *Frontiers in Neuroinformatics* 11. ISSN: 1662-5196. DOI: 10.3389/fninf.2017.00076. URL: <http://journal.frontiersin.org/article/10.3389/fninf.2017.00076/full> (visited on 04/21/2020).
- Pomerleau, Dean and Delip Rao (Dec. 2016). *Fake News Challenge*. Exploring how artificial intelligence technologies could be leveraged to combat fake news. URL: <https://www.fakenewschallenge.org/> (visited on 03/13/2020).
- Popat, Kashyap, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum (2017). “Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media”. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW ’17 Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 1003–1012. ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3055133. URL: <https://doi.org/10.1145/3041021.3055133>.
- Popat, Kashyap, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum (Oct. 11, 2018). “DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 22–32. URL: <https://www.aclweb.org/anthology/D18-1003>.
- Popper, Karl (2005). *Unended Quest: An Intellectual Autobiography*. Taylor and Francis. ISBN: 978-0-203-99425-2.
- Prost, Flavien, Nithum Thain, and Tolga Bolukbasi (Aug. 2019). “Debiasing Embeddings for Reduced Gender Bias in Text Classification”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, pp. 69–75. DOI: 10.18653/v1/W19-3810. URL: <https://www.aclweb.org/anthology/W19-3810>.
- Reimers, Nils, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych (Oct. 2014). “GermEval-2014: Nested Named Entity Recognition with Neural Networks”. In: *Workshop Proceedings of the 12th Edition of the KONVENS Conference*. Ed. by Gertrud Faaß and Josef Ruppenhofer. Hildesheim, Germany: Universitätsverlag Hildesheim, pp. 117–120.
- Reimers, Nils and Iryna Gurevych (Mar. 26, 2018). *Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions About Machine Learning Approaches*. arXiv: 1803.09578 [cs, stat]. URL: <http://arxiv.org/abs/1803.09578> (visited on 03/09/2020).
- Reimers, Nils, Benjamin Schiller, Tillman Beck, Johannes Daxenberger, and Iryna Gurevych (July 2019). “Classification and Clustering of Arguments with Contextualized Word Embeddings”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, to appear.
- Resnik, Philip and Jimmy Lin (Aug. 2010). “Evaluation of NLP Systems”. In: *The Handbook of Computational Linguistics and Natural Language Processing*. Blackwell Handbooks in Linguistics. Wiley-Blackwell, pp. 271–295. ISBN: 978-1-4051-5581-6. URL: https://www.cs.colorado.edu/~jbg/teaching/CMSC_773_2012/reading/evaluation.pdf.

- Rinott, Rutu, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim (Sept. 2015). “Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 440–450. URL: <http://aclweb.org/anthology/D15-1050>.
- Rubin, Victoria L., Yimin Chen, and Niall J. Conroy (2015). “Deception Detection for News: Three Types of Fakes: Deception Detection for News: Three Types of Fakes”. In: *Proceedings of the Association for Information Science and Technology* 52.1, pp. 1–4. ISSN: 23739231. DOI: 10.1002/pra2.2015.145052010083. URL: <http://doi.wiley.com/10.1002/pra2.2015.145052010083> (visited on 03/13/2020).
- Rus, Vasile and Mihai Lintean (June 2012). “A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics”. In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, pp. 157–162. URL: <https://www.aclweb.org/anthology/W12-2018>.
- Rutherford, Attapol and Santhawat Thanyawong (Aug. 2019). “Written on Leaves or in Stones?: Computational Evidence for the Era of Authorship of Old Thai Prose”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, pp. 81–85. URL: <https://www.aclweb.org/anthology/W19-4710> (visited on 07/31/2019).
- Salesky, Elizabeth, Matthias Sperber, and Alan W Black (July 2019). “Exploring Phoneme-Level Speech Representations for End-to-End Speech Translation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1835–1841. DOI: 10.18653/v1/P19-1179. URL: <https://www.aclweb.org/anthology/P19-1179>.
- Salmon, Wesley C. (1983). “Confirmation and Relevance”. In: *The Concept of Evidence*. Oxford Readings in Philosophy. New York, NY, USA: Oxford University Press, pp. 95–123. ISBN: 0-19-875062-5.
- Schaal, Stefan (1996). “Learning from Demonstration”. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS’96. Cambridge, MA, USA: MIT Press, pp. 1040–1046. URL: <http://dl.acm.org/citation.cfm?id=2998981.2999127> (visited on 04/21/2017).
- Schaffer, Simon (1992). “Self Evidence”. In: *Critical Inquiry* 18.2, pp. 327–362. DOI: 10.1086/448635. eprint: <https://doi.org/10.1086/448635>. URL: <https://doi.org/10.1086/448635>.
- Schlichtkrull, Michael and Anders Søgaard (Apr. 2017). “Cross-Lingual Dependency Parsing with Late Decoding for Truly Low-Resource Languages”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 220–229. URL: <https://www.aclweb.org/anthology/E17-1021>.
- Schulz, Claudia, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych (June 2018). “Multi-Task Learning for Argumentation Mining in Low-Resource Settings”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*

- (*Short Papers*). New Orleans, Louisiana: Association for Computational Linguistics, pp. 35–41. DOI: 10.18653/v1/N18-2006. URL: <https://www.aclweb.org/anthology/N18-2006>.
- Schulz, Claudia, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych (July 2019). “Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2761–2772. DOI: 10.18653/v1/P19-1265. URL: <https://www.aclweb.org/anthology/P19-1265>.
- Shardlow, Matthew, Riza Batista-Navarro, Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou (June 25, 2018). “Identification of Research Hypotheses and New Knowledge from Scientific Literature”. In: *BMC Medical Informatics and Decision Making* 18.1, p. 46. ISSN: 1472-6947. DOI: 10.1186/s12911-018-0639-1. URL: <https://doi.org/10.1186/s12911-018-0639-1> (visited on 09/11/2018).
- Shnarch, Eyal, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim (July 2018). “Will It Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 599–605. URL: <http://www.aclweb.org/anthology/P18-2095>.
- Sonoda, Sho and Noboru Murata (2019). “Transport Analysis of Infinitely Deep Neural Network”. In: *Journal of Machine Learning Research* 20.2, pp. 1–52. URL: <http://jmlr.org/papers/v20/16-243.html>.
- Stab, Christian, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych (June 2018a). “ArgumenText: Searching for Arguments in Heterogeneous Sources”. In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, USA: Association for Computational Linguistics.
- Stab, Christian and Iryna Gurevych (Aug. 2014). “Annotating Argument Components and Relations in Persuasive Essays.” In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1501–1510.
- (Sept. 2017). “Parsing Argumentation Structures in Persuasive Essays”. In: *Computational Linguistics* 43.3, pp. 619–659.
- Stab, Christian, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych (Oct. 11, 2018b). “Cross-Topic Argument Mining from Heterogeneous Sources”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3664–3674. URL: <https://www.aclweb.org/anthology/D18-1402>.
- Stahlhut, Chris (Aug. 2018). “Searching Arguments in German with ArgumenText”. In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. First Biennial Conference on Design of Experimental Search and Information Retrieval Systems. Vol. 2167. CEUR Workshop Proceedings. Bertinoro, Italy, p. 104. URL: <http://ceur-ws.org/Vol-2167/short7.pdf>.

- Stahlhut, Chris (Oct. 2019a). “Combating Disinformation via Interactive Evidence Detection”. In: *Proceedings of the First Conference on Truth and Trust Online*. London, UK, p. 9. URL: https://truthandtrustonline.files.wordpress.com/2019/09/paper_9.pdf.
- (Nov. 2019b). “Interactive Evidence Detection: Train State-of-the-Art Model out-of-Domain or Simple Model Interactively?” In: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Hong Kong, China: Association for Computational Linguistics, pp. 79–89. DOI: 10.18653/v1/D19-6613. URL: <https://www.aclweb.org/anthology/D19-6613>.
- Stahlhut, Chris, Nicolás Navarro-Guerrero, Cornelius Weber, and Stefan Wermter (2016). “Interaction in Reinforcement Learning Reduces the Need for Finely Tuned Hyperparameters in Complex Tasks”. In: *Kognitive Systeme 3*. DOI: <http://dx.doi.org/10.17185/dupublico/40718>.
- Stahlhut, Chris, Christian Stab, and Iryna Gurevych (Aug. 2018). “Pilot Experiments of Hypothesis Validation Through Evidence Detection for Historians”. In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. First Biennial Conference on Design of Experimental Search & Information Retrieval Systems. Vol. 2167. CEUR Workshop Proceedings. Bertinoro, Italy, pp. 83–89. URL: <http://ceur-ws.org/Vol-2167/paper7.pdf>.
- Stenetorp, Pontus, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun’ichi Tsujii (June 2011). “BioNLP Shared Task 2011: Supporting Resources”. In: *Proceedings of BioNLP Shared Task 2011 Workshop*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 112–120. URL: <http://www.aclweb.org/anthology/W11-1816>.
- Strauss, Anselm and Juliet Corbin (1994). “Grounded Theory Methodology: An Overview.” In: *Handbook of Qualitative Research*. Thousand Oaks, CA, US: Sage Publications, Inc, pp. 273–285. ISBN: 0-8039-4679-1 (Hardcover).
- Sweeney, Chris and Maryam Najafian (July 2019). “A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1662–1667. DOI: 10.18653/v1/P19-1162. URL: <https://www.aclweb.org/anthology/P19-1162>.
- Thomaz, Andrea L. and Cynthia Breazeal (Apr. 1, 2008). “Teachable Robots: Understanding Human Teaching Behavior to Build More Effective Robot Learners”. In: *Artificial Intelligence* 172.6, pp. 716–737. ISSN: 0004-3702. DOI: 10.1016/j.artint.2007.09.009. URL: <http://www.sciencedirect.com/science/article/pii/S000437020700135X> (visited on 02/24/2017).
- Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal (June 2018a). “FEVER: A Large-Scale Dataset for Fact Extraction and VERification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: <https://www.aclweb.org/anthology/N18-1074> (visited on 07/24/2019).

- Thorne, James, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal (Nov. 2018b). “The Fact Extraction and VERification (FEVER) Shared Task”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, pp. 1–9. DOI: 10.18653/v1/W18-5501. URL: <https://www.aclweb.org/anthology/W18-5501> (visited on 07/24/2019).
- Trenta, Antonio, Anthony Hunter, and Sebastian Riedel (2015). “Extraction of Evidence Tables from Abstracts of Randomized Clinical Trials Using a Maximum Entropy Classifier and Global Constraints”. In: *CoRR* abs/1509.05209. URL: <http://arxiv.org/abs/1509.05209>.
- Visser, Jacky, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed (Mar. 2020). “Argumentation in the 2016 US Presidential Elections: Annotated Corpora of Television Debates and Social Media Reaction”. In: *Language Resources and Evaluation* 54.1, pp. 123–154. ISSN: 1574-020X, 1574-0218. DOI: 10.1007/s10579-019-09446-8. URL: <http://link.springer.com/10.1007/s10579-019-09446-8> (visited on 04/25/2020).
- Vlachos, Andreas and Sebastian Riedel (June 2014). “Fact Checking: Task Definition and Dataset Construction”. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, pp. 18–22. DOI: 10.3115/v1/W14-2508. URL: <https://www.aclweb.org/anthology/W14-2508>.
- Wachsmuth, Henning, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein (Sept. 2017). “Building an Argument Search Engine for the Web”. In: *Proceedings of the 4th Workshop on Argument Mining*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 49–59. DOI: 10.18653/v1/W17-5106. URL: <https://www.aclweb.org/anthology/W17-5106>.
- Wachsmuth, Henning, Benno Stein, and Yamen Ajjour (Apr. 2017). “PageRank for Argument Relevance”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 1117–1127. URL: <http://www.aclweb.org/anthology/E17-1105>.
- Wang, William Yang (July 2017). ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2017. Vancouver, Canada: Association for Computational Linguistics, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: <https://www.aclweb.org/anthology/P17-2067> (visited on 07/24/2019).
- Wolff, Jonathan (2017). “Karl Marx”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/win2017/entries/marx/>.
- Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy (June 2016). “Hierarchical Attention Networks for Document Classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association

- for Computational Linguistics, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: <http://www.aclweb.org/anthology/N16-1174>.
- Yimam, Seid Muhie, Steffen Remus, Alexander Panchenko, Andreas Holzinger, and Chris Biemann (2017). “Entity-Centric Information Access with the Human-in-the-Loop for the Biomedical Domains”. In: *Biomedical NLP Workshop*. Associated with RANLP 2017. Varna, Bulgaria. URL: <https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2017-yimametal-bionlpatranlp.pdf> (visited on 09/18/2017).
- Yin, Wenpeng and Dan Roth (Oct. 11, 2018). “TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 105–114. URL: <https://www.aclweb.org/anthology/D18-1010>.
- Zhang, Yi, Zachary Ives, and Dan Roth (July 2019). “Evidence-Based Trustworthiness”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 413–423. URL: <https://www.aclweb.org/anthology/P19-1040> (visited on 07/31/2019).
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (June 2019). “Gender Bias in Contextualized Word Embeddings”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 629–634. DOI: 10.18653/v1/N19-1064. URL: <https://www.aclweb.org/anthology/N19-1064>.
- Zoghi, Masrour, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen (July 17, 2017). “Online Learning to Rank in Stochastic Click Models”. In: *PMLR*. International Conference on Machine Learning, pp. 4199–4208. URL: <http://proceedings.mlr.press/v70/zoghi17a.html> (visited on 10/16/2017).

Appendix A

Consent Form and Task Description for User Study

A.1 Consent Form



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Aufklärungsbogen & Einverständniserklärung

Die Richtlinien der Deutschen Forschungsgemeinschaft (DFG) sehen vor, dass sich die Teilnehmer_innen an empirischen Studien mit ihrer Unterschrift explizit und nachvollziehbar einverstanden erklären, um damit zu dokumentieren, dass sie freiwillig an unserer Forschung teilnehmen.

Aus diesem Grund möchten wir Sie bitten, die nachfolgenden Erläuterungen zu lesen und vorliegende Einverständniserklärung zu unterzeichnen, bevor Sie an unserer Studie teilnehmen.

Gegenstand der Studie/des Experiments

Ziel dieser Studie ist zu erfahren wie Geisteswissenschaftler_innen mit einem Softwarewerkzeug arbeiten in dem sie Belege markieren und Hypothesen zuordnen können.

Ablauf der Studie/des Experiments

Zeitplan [min]	Tätigkeit
0 – 5	Diese Einverständniserklärung
5 – 10	Einführung in die Software
10 – 15	Fragen zur Aufgabenstellung
15 – 75	Bearbeitung der Aufgabe
75 – 80	Ausfüllen des Fragebogens
80 – 90	Diskussion der Ergebnisse der Aufgabenstellung

Dauer

Die Teilnahme an der Studie/an dem Experiment wird voraussichtlich 90 Minuten in Anspruch nehmen.

Möglicher Nutzen der Studie/des Experiments

Der mögliche Nutzen dieser Studie ist eine bessere Unterstützung geisteswissenschaftlicher Arbeit mit Texten mit Hilfe digitaler Methoden, insbesondere Methoden des Information Retrievals und Künstlichen Intelligenz.

Mit der Teilnahme verbundene Erfahrungen/Risiken

Die Teilnehmer_innen an dieser Studie werden keinem Risiko ausgesetzt, das über die Risiken des alltäglichen Lebens hinausgeht.

Je nach Studie: Der genaue Zweck dieser Untersuchung kann den Teilnehmer und Teilnehmerinnen erst am Ende der Studie ausführlich und vollständig dargestellt werden, da die Gültigkeit der Ergebnisse ansonsten beeinflusst werden könnte. Mit der Unterzeichnung der Einverständniserklärung erklären Sie sich damit einverstanden, erst nach Beendigung der Studie über deren genauen Zweck informiert zu werden.

Vertraulichkeit

Ihre Daten sind selbstverständlich vertraulich und werden nur in anonymisierter Form genutzt. Demographische Angaben wie Alter oder Geschlecht lassen keinen eindeutigen Schluss auf Ihre Person zu. Zu keinem Zeitpunkt im Rahmen der jeweiligen Untersuchung werden wir Sie bitten, Ihren Namen oder andere eindeutige Informationen zu nennen.

Die Speicherung erfolgt in einer Form, die keinen Rückschluss auf Ihre Person zulässt, das heisst die Daten werden anonymisiert. Diese Einverständniserklärung wird getrennt von den anderen Versuchsmaterialien & Unterlagen aufbewahrt.

Freiwilligkeit

Ihre Teilnahme an dieser Untersuchung ist freiwillig. Es steht Ihnen zu jedem Zeitpunkt dieser Studie frei, Ihre Teilnahme abubrechen und damit diese Einwilligung zurückziehen, ohne dass Ihnen daraus Nachteile entstehen. Wenn Sie die Teilnahme abbrechen, werden keine Daten von Ihnen gespeichert und alle bisher vorliegenden Daten zu Ihrer Person vernichtet.

Einverständnis

Ich habe die Erläuterungen zur Studie/zum Experiment gelesen und bin damit einverstanden, an der genannten Studie/dem genannten Experiment teilzunehmen.

Ich erkläre mich einverstanden, dass die im Rahmen der Studie/des Experiments Daten erhobenen Daten zu wissenschaftlichen Zwecken ausgewertet und in anonymisierter Form gespeichert und veröffentlicht werden. Ich bin mir darüber bewusst, dass meine Teilnahme freiwillig erfolgt und ich den Versuch jederzeit und ohne die Angabe von Gründen abbrechen kann.

Datum

Name

Unterschrift

Bei Fragen, Anregungen oder Beschwerden können Sie sich gerne an den Versuchsleiter wenden:

Chris Stahlhut

Email: stahlhut@kritis.tu-darmstadt.de

A.2 Task Descriptions for User Studies and Feedback forms

Vergleich des politischen Diskurses nach nuklearen Störfällen

1. Zugangsdaten

Username: _____

Password: _____

Der Login Dialog sollte bereits aufgerufen sein. Sofern dies nicht der Fall ist, öffnen Sie bitte Chrome in der Menüleiste am unteren Bildschirmrand und navigieren zu:

<http://argsim-study.ukp.informatik.tu-darmstadt.de>

und loggen sich mit den hier angegebenen Zugangsdaten ein.

Chrome können Sie auf den Desktop Rechnern anhand des Icons



erkennen.

2. Aufgabenstellung

Die Störfälle der Atomkraftwerke in Tschernobyl und Fukushima haben gezeigt, welche Konsequenzen die Verwendung nuklearer Energie haben kann. Im Rahmen dieser Übung sollen Hypothesen bezüglich der Unterschiede und Gemeinsamkeiten des politischen Diskurses nach solchen Ereignissen erarbeitet und mit Evidenzen belegt werden. Die folgenden Fragen können Ihnen als Leitfaden dienen:

1. Welche Themen werden nach nur einem Ereignis kontrovers diskutiert?
2. Welche Themen werden ereignisübergreifend kontrovers diskutiert?
3. Was für Beziehungen werden zwischen den beiden Ereignissen hergestellt?

In der bereitgestellten Anwendung können Sie hierzu Evidenzen in Quelldokumenten markieren. Bitte markieren Sie dafür möglichst ganze Zeilen/Sätze. Ihre Hypothesen können Sie in dem „Hypothesen/Evidenzen Tab“ formulieren, ändern und ergänzen, und mit den zuvor markierten Evidenzen belegen.

3. Feedback

Hat Ihnen das System bei der Konkretisierung Ihrer Hypothesen geholfen?

Ja ☐ Nein ☐

Haben Sie Ihre Hypothesen während der Übung geändert?

Ja ☐ Nein ☐

Welche Arbeitsweise beschreibt Ihr Vorgehen am besten?

Zunächst Evidenzen finden, dann Hypothesen definieren

☐

Zunächst Hypothesen definieren, dann Evidenzen finden

☐

Andere (bitte Beschreiben)

☐

Konnten Sie durch die Übung neue Erkenntnisse gewinnen?

Ja ☐ Nein ☐

Können Sie sich vorstellen dieses System für weitere Rechercheaufgaben im Rahmen Ihres Studiums zu nutzen?
(Wenn nicht, erläutern Sie bitte warum)

Ja ☐ Nein ☐

Was gefiel Ihnen an der Anwendung?

Was gefiel Ihnen **nicht** an der Anwendung?

Welche Änderungen wünschen Sie sich für die Zukunft dieser Anwendung?

Appendix B

Source Listing

```
#!/usr/bin/env groovy
@Grab('org.dkpro.script:dkpro-script-groovy:0.1.0')
@groovy.transform.BaseScript org.dkpro.script.groovy.DKProCoreScript baseScript
read 'Text' language 'de' from '$DATA_FOLDER/*.txt'
apply 'OpenNlpSegmenter'
write 'Xmi' to 'results' params([typeSystemFile: '/dev/null'])
```

Listing B.1: The DKPro Script to conduct pre-processing of plain text files before imporing them into EDoHa. We do not need the type system description and therefore discard it.

```
def annotate_sentences(sentences, evidential_sentences):
    annotations = [False] * len(sentences)
    found = set()
    for i, annotatable in enumerate(sentences):
        for j, evidence in enumerate(evidential_sentences):
            if evidence.lower() in annotatable.lower():
                annotations[i] = True
                found.add(evidence)
    return annotations
```

Listing B.2: The algorithm we used to convert lists of evidence to evidence-annotated documents.

```
def extend_negative(links, hypotheses, evidences):
    hyp2idx = {hyp: idx for idx, hyp in enumerate(hypotheses)}
    ev2idx = {ev: idx for idx, ev in enumerate(evidences)}

    adjacency = np.zeros((len(hypotheses), len(evidences)))
    for link in links:
        adjacency[hyp2idx[link[0]], ev2idx[link[1]]] = 1

    random_links = list()
    for (hyp, _) in links:
        hyp_idx = hyp2idx[hyp]
        evidence_indices = adjacency[hyp_idx]
        unlinked_evidences = np.where(evidence_indices == 0)[0]
        if not len(unlinked_evidences): break
        random_evidence = np.random.choice(unlinked_evidences)
        link_with_text = (hyp, evidences[random_evidence])
        random_links.append(link_with_text)
        # make sure this connection cannot be drawn again
        adjacency[hyp_idx][random_evidence] = 1

    data = np.concatenate((links, random_links))
    labels = np.concatenate((["link"] * len(links), ["no-link"] * len(random_links)))
    complete = np.concatenate((data, np.expand_dims(labels, axis=1)), axis=1)
```

```
return complete
```

Listing B.3: The algorithm we used to create non-links in our testing datasets.

Appendix C

Hyper-Parameter Optimisation for Machine Learning for Evidence Linking

Table C.1: Results of the individual hyper-parameter configurations of the bilstmAtt model. The labels for the different hyper-parameters are *lr* for the learning rate, *ep* for the number of epochs, and *bs* for the batch size.

Model	Group1	Group2	Group3	Group4
lr=0.002, ep=40, bs=4	0.554	0.553	0.502	0.493
lr=0.002, ep=40, bs=2	0.549	0.544	0.489	0.490
lr=0.0001, ep=5, bs=4	0.434	0.396	0.413	0.466
lr=0.0005, ep=20, bs=4	0.545	0.568	0.507	0.482
lr=0.0005, ep=5, bs=16	0.442	0.398	0.448	0.487
lr=0.0001, ep=20, bs=1	0.535	0.542	0.485	0.484
lr=0.001, ep=10, bs=2	0.530	0.538	0.487	0.482
lr=0.0005, ep=40, bs=16	0.559	0.549	0.513	0.485
lr=0.0005, ep=20, bs=16	0.562	0.557	0.521	0.487
lr=0.0001, ep=10, bs=2	0.438	0.405	0.484	0.494
lr=0.0005, ep=5, bs=4	0.449	0.445	0.476	0.487
lr=0.0005, ep=20, bs=1	0.548	0.539	0.498	0.475
lr=0.002, ep=20, bs=1	0.556	0.514	0.482	0.488
lr=0.001, ep=5, bs=2	0.512	0.522	0.487	0.484
lr=0.0001, ep=40, bs=4	0.542	0.539	0.512	0.498
lr=0.002, ep=20, bs=4	0.534	0.552	0.492	0.479
lr=0.0001, ep=5, bs=2	0.440	0.392	0.405	0.470
lr=0.0005, ep=10, bs=16	0.462	0.450	0.489	0.488
lr=0.0001, ep=40, bs=16	0.453	0.467	0.499	0.492
lr=0.001, ep=5, bs=4	0.536	0.531	0.497	0.487
lr=0.002, ep=5, bs=4	0.535	0.520	0.477	0.487
lr=0.0001, ep=10, bs=4	0.445	0.398	0.447	0.489
lr=0.0005, ep=40, bs=2	0.515	0.536	0.487	0.487
lr=0.002, ep=5, bs=2	0.528	0.505	0.493	0.491
lr=0.002, ep=20, bs=2	0.554	0.550	0.500	0.480
lr=0.002, ep=20, bs=16	0.523	0.547	0.500	0.483
lr=0.001, ep=40, bs=1	0.557	0.556	0.497	0.515
lr=0.0005, ep=10, bs=2	0.527	0.528	0.486	0.482
lr=0.001, ep=20, bs=1	0.551	0.563	0.488	0.485
lr=0.002, ep=5, bs=1	0.512	0.509	0.479	0.475
lr=0.0001, ep=20, bs=16	0.429	0.390	0.446	0.492
lr=0.002, ep=5, bs=16	0.500	0.504	0.482	0.487
lr=0.002, ep=40, bs=1	0.545	0.511	0.467	0.499
lr=0.0005, ep=5, bs=2	0.485	0.511	0.485	0.483
lr=0.001, ep=40, bs=2	0.552	0.554	0.495	0.497
lr=0.0001, ep=20, bs=2	0.509	0.532	0.495	0.499
lr=0.001, ep=5, bs=16	0.451	0.434	0.469	0.482
lr=0.001, ep=40, bs=4	0.529	0.558	0.509	0.486
lr=0.0005, ep=40, bs=1	0.547	0.549	0.488	0.503
lr=0.001, ep=10, bs=1	0.537	0.540	0.494	0.486
lr=0.0001, ep=10, bs=1	0.447	0.389	0.431	0.487
lr=0.001, ep=20, bs=4	0.539	0.536	0.494	0.476
lr=0.0001, ep=5, bs=1	0.423	0.383	0.372	0.414
lr=0.002, ep=10, bs=16	0.512	0.558	0.494	0.481
lr=0.002, ep=10, bs=1	0.516	0.505	0.487	0.482
lr=0.001, ep=10, bs=4	0.559	0.552	0.492	0.479
lr=0.0005, ep=10, bs=1	0.537	0.542	0.505	0.481
lr=0.0001, ep=40, bs=1	0.549	0.549	0.484	0.484
lr=0.002, ep=40, bs=16	0.524	0.542	0.501	0.492
lr=0.0001, ep=20, bs=4	0.460	0.483	0.486	0.495
lr=0.001, ep=20, bs=16	0.514	0.546	0.501	0.487
lr=0.0001, ep=10, bs=16	0.444	0.389	0.399	0.466
lr=0.001, ep=5, bs=1	0.543	0.493	0.486	0.489
lr=0.002, ep=10, bs=4	0.540	0.531	0.487	0.482
lr=0.001, ep=40, bs=16	0.530	0.539	0.503	0.480
lr=0.002, ep=10, bs=2	0.522	0.513	0.494	0.467
lr=0.0001, ep=5, bs=16	0.450	0.382	0.385	0.453
lr=0.0005, ep=10, bs=4	0.538	0.538	0.502	0.487
lr=0.0005, ep=40, bs=4	0.535	0.566	0.509	0.483
lr=0.001, ep=20, bs=2	0.557	0.539	0.501	0.482
lr=0.0005, ep=5, bs=1	0.466	0.472	0.491	0.479
lr=0.001, ep=10, bs=16	0.523	0.549	0.495	0.492
lr=0.0001, ep=40, bs=2	0.539	0.546	0.505	0.483
lr=0.0005, ep=20, bs=2	0.493	0.535	0.503	0.479

Table C.2: Results of the individual hyper-parameter configurations of the `bilstmAttfine` model for the nuclear energy dataset. The different hyper-parameters are *lr* for the learning rate, *ilr* for the learning rate in the initial fine-tuning phase, *e* for the number of epochs, *ie* for the number of epochs in the initial fine-tuning phase, and *bs* for the batch size.

Model	Group1	Group2	Group3	Group4
lr=0.0001, ilr=0.0005, e=0, ie=1, bs=2	0.510	0.430	0.459	0.472
lr=0.0001, ilr=0.0001, e=1, ie=5, bs=2	0.486	0.454	0.446	0.478
lr=0.0005, ilr=0.002, e=10, ie=5, bs=2	0.371	0.296	0.339	0.457
lr=0.0005, ilr=0.002, e=5, ie=0, bs=1	0.371	0.328	0.321	0.436
lr=0.0005, ilr=0.002, e=1, ie=0, bs=16	0.495	0.452	0.477	0.474
lr=0.0001, ilr=0.0005, e=1, ie=10, bs=2	0.495	0.426	0.313	0.437
lr=0.001, ilr=0.001, e=1, ie=2, bs=16	0.455	0.416	0.398	0.452
lr=0.002, ilr=0.001, e=0, ie=1, bs=16	0.510	0.430	0.466	0.477
lr=0.001, ilr=0.0005, e=5, ie=0, bs=1	0.358	0.305	0.334	0.429
lr=0.001, ilr=0.0001, e=0, ie=10, bs=2	0.510	0.430	0.419	0.475
lr=0.0001, ilr=0.002, e=2, ie=2, bs=1	0.370	0.343	0.342	0.422
lr=0.0005, ilr=0.0005, e=10, ie=5, bs=16	0.377	0.315	0.317	0.429
lr=0.0001, ilr=0.002, e=0, ie=10, bs=16	0.510	0.430	0.322	0.425
lr=0.0001, ilr=0.0001, e=1, ie=0, bs=2	0.510	0.446	0.484	0.491
lr=0.0005, ilr=0.002, e=10, ie=2, bs=2	0.371	0.296	0.339	0.457
lr=0.0005, ilr=0.002, e=1, ie=10, bs=2	0.379	0.369	0.305	0.435
lr=0.0001, ilr=0.0005, e=5, ie=10, bs=2	0.376	0.321	0.312	0.433
lr=0.002, ilr=0.0001, e=2, ie=2, bs=2	0.347	0.324	0.333	0.408
lr=0.001, ilr=0.0001, e=10, ie=5, bs=2	0.380	0.320	0.364	0.464
lr=0.0001, ilr=0.002, e=5, ie=10, bs=16	0.382	0.322	0.317	0.430
lr=0.0001, ilr=0.002, e=2, ie=2, bs=4	0.378	0.390	0.363	0.432
lr=0.0005, ilr=0.001, e=0, ie=1, bs=1	0.510	0.430	0.393	0.424
lr=0.0001, ilr=0.0001, e=2, ie=10, bs=16	0.520	0.441	0.466	0.485
lr=0.0001, ilr=0.0005, e=5, ie=5, bs=1	0.376	0.321	0.315	0.437
lr=0.001, ilr=0.002, e=5, ie=5, bs=4	0.365	0.304	0.328	0.445
lr=0.002, ilr=0.002, e=2, ie=0, bs=4	0.365	0.315	0.331	0.406
lr=0.0001, ilr=0.001, e=0, ie=1, bs=16	0.510	0.430	0.466	0.477
lr=0.001, ilr=0.002, e=0, ie=10, bs=4	0.510	0.430	0.324	0.426
lr=0.002, ilr=0.002, e=1, ie=10, bs=4	0.371	0.362	0.333	0.434
lr=0.0001, ilr=0.002, e=0, ie=5, bs=2	0.510	0.430	0.317	0.418
lr=0.001, ilr=0.0001, e=1, ie=2, bs=16	0.486	0.439	0.408	0.466
lr=0.0001, ilr=0.002, e=1, ie=2, bs=16	0.465	0.414	0.392	0.457
lr=0.002, ilr=0.002, e=0, ie=2, bs=16	0.510	0.430	0.396	0.460
lr=0.001, ilr=0.0001, e=5, ie=2, bs=2	0.362	0.314	0.316	0.437
lr=0.0001, ilr=0.001, e=10, ie=5, bs=2	0.351	0.308	0.317	0.426
lr=0.0001, ilr=0.0005, e=1, ie=10, bs=16	0.494	0.458	0.371	0.449
lr=0.002, ilr=0.002, e=0, ie=10, bs=16	0.510	0.430	0.322	0.425
lr=0.001, ilr=0.0005, e=10, ie=10, bs=1	0.375	0.316	0.383	0.490
lr=0.001, ilr=0.0005, e=5, ie=0, bs=16	0.380	0.350	0.329	0.430
lr=0.002, ilr=0.0005, e=5, ie=5, bs=4	0.349	0.310	0.319	0.444
lr=0.0001, ilr=0.0001, e=0, ie=5, bs=1	0.510	0.430	0.453	0.481
lr=0.001, ilr=0.0001, e=5, ie=1, bs=4	0.372	0.310	0.325	0.432
lr=0.002, ilr=0.0005, e=5, ie=2, bs=4	0.349	0.310	0.306	0.438
lr=0.0001, ilr=0.0005, e=5, ie=2, bs=2	0.376	0.321	0.355	0.437
lr=0.0005, ilr=0.001, e=5, ie=2, bs=4	0.363	0.311	0.322	0.435
lr=0.0001, ilr=0.001, e=10, ie=5, bs=1	0.381	0.313	0.320	0.429
lr=0.0001, ilr=0.0005, e=1, ie=1, bs=16	0.494	0.458	0.488	0.488
lr=0.002, ilr=0.001, e=2, ie=5, bs=2	0.369	0.335	0.325	0.429
lr=0.001, ilr=0.001, e=0, ie=5, bs=4	0.510	0.430	0.328	0.433
lr=0.001, ilr=0.001, e=0, ie=5, bs=1	0.510	0.430	0.323	0.425
lr=0.002, ilr=0.0001, e=2, ie=0, bs=2	0.376	0.313	0.328	0.410
lr=0.0001, ilr=0.002, e=10, ie=2, bs=1	0.387	0.303	0.318	0.421
lr=0.001, ilr=0.001, e=1, ie=1, bs=2	0.399	0.360	0.348	0.404
lr=0.0005, ilr=0.001, e=1, ie=2, bs=1	0.394	0.348	0.340	0.420
lr=0.002, ilr=0.0005, e=5, ie=0, bs=16	0.360	0.311	0.317	0.419
lr=0.0001, ilr=0.0001, e=2, ie=1, bs=2	0.479	0.449	0.461	0.480
lr=0.0001, ilr=0.0005, e=10, ie=2, bs=1	0.359	0.320	0.317	0.458
lr=0.0005, ilr=0.0005, e=1, ie=2, bs=4	0.460	0.426	0.392	0.443
lr=0.0005, ilr=0.0001, e=1, ie=1, bs=4	0.463	0.445	0.432	0.475
lr=0.002, ilr=0.001, e=0, ie=5, bs=16	0.510	0.430	0.392	0.456

Table C.3: Results of the individual hyper-parameter configurations of the `bilstmAttfine` model for the forest dieback dataset. The different hyper-parameters are lr for the learning rate, ilr for the learning rate in the initial fine-tuning phase, e for the number of epochs, ie for the number of epochs in the initial fine-tuning phase, and bs for the batch size.

Model	Group1	Group2	Group3	Group4
$lr=0.0001, ilr=0.0005, e=0, ie=1, bs=2$	0.484	0.491	0.433	0.503
$lr=0.0001, ilr=0.0001, e=1, ie=5, bs=2$	0.488	0.476	0.431	0.501
$lr=0.0005, ilr=0.002, e=10, ie=5, bs=2$	0.337	0.394	0.399	0.506
$lr=0.0005, ilr=0.002, e=5, ie=0, bs=1$	0.369	0.395	0.374	0.496
$lr=0.0005, ilr=0.002, e=1, ie=0, bs=16$	0.491	0.497	0.452	0.508
$lr=0.0001, ilr=0.0005, e=1, ie=10, bs=2$	0.450	0.396	0.375	0.494
$lr=0.001, ilr=0.001, e=1, ie=2, bs=16$	0.423	0.438	0.423	0.475
$lr=0.002, ilr=0.001, e=0, ie=1, bs=16$	0.500	0.510	0.437	0.511
$lr=0.001, ilr=0.0005, e=5, ie=0, bs=1$	0.337	0.383	0.366	0.499
$lr=0.001, ilr=0.0001, e=0, ie=10, bs=2$	0.477	0.454	0.417	0.503
$lr=0.0001, ilr=0.002, e=2, ie=2, bs=1$	0.332	0.396	0.383	0.484
$lr=0.0005, ilr=0.0005, e=10, ie=5, bs=16$	0.347	0.401	0.368	0.487
$lr=0.0001, ilr=0.002, e=0, ie=10, bs=16$	0.477	0.416	0.398	0.483
$lr=0.0001, ilr=0.0001, e=1, ie=0, bs=2$	0.484	0.501	0.446	0.509
$lr=0.0005, ilr=0.002, e=10, ie=2, bs=2$	0.333	0.395	0.381	0.500
$lr=0.0005, ilr=0.002, e=1, ie=10, bs=2$	0.326	0.377	0.370	0.507
$lr=0.0001, ilr=0.0005, e=5, ie=10, bs=2$	0.352	0.398	0.382	0.496
$lr=0.002, ilr=0.0001, e=2, ie=2, bs=2$	0.310	0.373	0.371	0.454
$lr=0.001, ilr=0.0001, e=10, ie=5, bs=2$	0.374	0.380	0.381	0.514
$lr=0.0001, ilr=0.002, e=5, ie=10, bs=16$	0.357	0.396	0.391	0.488
$lr=0.0001, ilr=0.002, e=2, ie=2, bs=4$	0.365	0.424	0.416	0.482
$lr=0.0005, ilr=0.001, e=0, ie=1, bs=1$	0.410	0.452	0.452	0.473
$lr=0.0001, ilr=0.0001, e=2, ie=10, bs=16$	0.497	0.504	0.436	0.517
$lr=0.0001, ilr=0.0005, e=5, ie=5, bs=1$	0.351	0.398	0.375	0.496
$lr=0.001, ilr=0.002, e=5, ie=5, bs=4$	0.333	0.390	0.385	0.504
$lr=0.002, ilr=0.002, e=2, ie=0, bs=4$	0.361	0.391	0.380	0.473
$lr=0.0001, ilr=0.001, e=0, ie=1, bs=16$	0.500	0.510	0.437	0.511
$lr=0.001, ilr=0.002, e=0, ie=10, bs=4$	0.477	0.505	0.396	0.496
$lr=0.002, ilr=0.002, e=1, ie=10, bs=4$	0.336	0.380	0.390	0.492
$lr=0.0001, ilr=0.002, e=0, ie=5, bs=2$	0.333	0.397	0.387	0.495
$lr=0.001, ilr=0.0001, e=1, ie=2, bs=16$	0.498	0.470	0.426	0.486
$lr=0.0001, ilr=0.002, e=1, ie=2, bs=16$	0.405	0.444	0.435	0.483
$lr=0.002, ilr=0.002, e=0, ie=2, bs=16$	0.409	0.444	0.434	0.473
$lr=0.001, ilr=0.0001, e=5, ie=2, bs=2$	0.323	0.393	0.370	0.499
$lr=0.0001, ilr=0.001, e=10, ie=5, bs=2$	0.331	0.386	0.383	0.489
$lr=0.0001, ilr=0.0005, e=1, ie=10, bs=16$	0.498	0.432	0.395	0.503
$lr=0.002, ilr=0.002, e=0, ie=10, bs=16$	0.477	0.505	0.398	0.483
$lr=0.001, ilr=0.0005, e=10, ie=10, bs=1$	0.392	0.399	0.379	0.532
$lr=0.001, ilr=0.0005, e=5, ie=0, bs=16$	0.371	0.400	0.385	0.479
$lr=0.002, ilr=0.0005, e=5, ie=5, bs=4$	0.345	0.382	0.363	0.503
$lr=0.0001, ilr=0.0001, e=0, ie=5, bs=1$	0.483	0.476	0.423	0.514
$lr=0.001, ilr=0.0001, e=5, ie=1, bs=4$	0.308	0.404	0.381	0.475
$lr=0.002, ilr=0.0005, e=5, ie=2, bs=4$	0.300	0.388	0.366	0.498
$lr=0.0001, ilr=0.0005, e=5, ie=2, bs=2$	0.392	0.419	0.386	0.487
$lr=0.0005, ilr=0.001, e=5, ie=2, bs=4$	0.350	0.405	0.369	0.487
$lr=0.0001, ilr=0.001, e=5, ie=5, bs=1$	0.341	0.394	0.393	0.503
$lr=0.0001, ilr=0.0005, e=1, ie=1, bs=16$	0.498	0.505	0.448	0.511
$lr=0.002, ilr=0.001, e=2, ie=5, bs=2$	0.362	0.404	0.379	0.481
$lr=0.001, ilr=0.001, e=0, ie=5, bs=4$	0.354	0.401	0.378	0.493
$lr=0.001, ilr=0.001, e=0, ie=5, bs=1$	0.343	0.412	0.395	0.500
$lr=0.002, ilr=0.0001, e=2, ie=0, bs=2$	0.325	0.394	0.367	0.462
$lr=0.0001, ilr=0.002, e=10, ie=2, bs=1$	0.319	0.396	0.381	0.494
$lr=0.001, ilr=0.001, e=1, ie=1, bs=2$	0.355	0.426	0.392	0.447
$lr=0.0005, ilr=0.001, e=1, ie=2, bs=1$	0.337	0.413	0.383	0.466
$lr=0.002, ilr=0.0005, e=5, ie=0, bs=16$	0.327	0.391	0.358	0.476
$lr=0.0001, ilr=0.0001, e=2, ie=1, bs=2$	0.495	0.486	0.441	0.509
$lr=0.0001, ilr=0.0005, e=10, ie=2, bs=1$	0.365	0.385	0.382	0.492
$lr=0.0005, ilr=0.0005, e=1, ie=2, bs=4$	0.426	0.442	0.404	0.470
$lr=0.0005, ilr=0.0001, e=1, ie=1, bs=4$	0.475	0.489	0.425	0.484
$lr=0.002, ilr=0.001, e=0, ie=5, bs=16$	0.477	0.505	0.416	0.496

List of Figures

1.1	The five identified NLP tasks necessary to automatically validate hypotheses in a pipeline model. The tasks with the gray striped background are not considered in this thesis.	3
1.2	Structure of the remainder of this thesis and the dependencies between the chapters. The circled numbers show the relations to the contributions of this thesis.	8
2.1	Illustration of the differences between the hypotheses that all emeralds are green and that all emeralds are grue.	14
2.2	Illustration of a dataset split into training, develop, and testing data. . . .	15
2.3	Illustration of a three-fold cross validation.	16
3.1	Different approaches to ED. a shows a ranking approach, whereas b shows a classification approach with hypothesis, and c shows a classification approach without access to the hypothesis.	26
3.2	The different approaches towards CV. b illustrates a model with task exactly one task, whereas c shows a model that performs ED and CV jointly while a shows a pipeline of different pre-trained models.	30
4.1	Document view of EDoHa. The numbers refer to the individual components described in the main text.	43
4.2	Evidence linking view of EDoHa. The numbers refer to the individual components described in the main text.	44
4.3	The number of evidence annotations (+), evidence/hypothesis links (x) for validation, and hypotheses (•) over time.	47
4.4	Activity of the users in documents over time. Each row represents one document sorted alphabetically from top to bottom. Each vertical bar represents an event, either the creation or deletion of an annotation, in the particular document.	49
4.5	The growth of hypotheses for each user over time. Each layer represents a hypothesis, and the height represents the amount of evidence linked to it. User2 and User3 did not create any link between a piece of evidence and a hypothesis.	50
5.1	Overview of how the user-dependent, in-domain data and user-independent, out-of-domain data relate to models and the different sub-questions.	62
5.2	The relation between the out-of-domain and in-domain datasets and different training setups.	68
5.3	Histogram detailing the variation in the lengths of evidence in the ED-ACL-2014 dataset.	70

5.4	Histogram detailing the variation in the lengths of evidence in the ED-EMNLP-2015 dataset.	70
5.5	Illustration of the fine-tuning for the $\text{bilstm}_{\text{fine}}$ model. The layer without background colour is added in the second step, and the nodes with a patterned background are trained in this step.	74
5.6	Sequence diagram of the simulated user interactively creating training data, training ED models, and correcting the predictions. The user picks an unannotated document and labels the evidential sentences. After processing the document, it is added to the training data for a newly trained model. Afterwards, the user picks the next document, which contains suggestions from the model.	81
6.1	Architecture of the bilstmAtt network. The individual tokens of the hypothesis (h_0, h_1, h_2) are encoded into the individual hypothesis vector (h). The tokens of the sentence (s_0, s_1, s_2, s_3) are encoded into contextual token embeddings (cs_0, cs_1, cs_2, cs_3) and then weighed with hypothesis attention. The attention-weighted contextual token embeddings ($acs_0, acs_1, acs_2, acs_3$) are then fed into a dense layer for classification.	90
6.2	Transfer learning approach for the $\text{bilstmAtt}_{\text{fine}}$. The grey nodes are pre-trained, and the node without background colour is the newly added classifier. The nodes with the patterned background are trained or fine-tuned in the current step.	90
7.1	Screenshot of EDoHa's document view with the ability to train an ED model. The numbers refer to the additions made to incorporate suggestions.	101
7.2	Screenshot of the evidence linking view of EDoHa with the ability to train an EL model. The numbers refer to the additions made to incorporate suggestions.	101
7.3	After reading the first two documents d_1 and d_2 , the user clicks on Train Model to train the first model m_1 . This is then used to suggest evidence on the third document d_3 , which the user corrects and after clicking again on Train Model , the system trains the second model m_2 on the training documents d_1, d_2 , and d_3	102
7.4	The user first creates two links and clicks the Train Model button. Then EDoHa generates two weakly labelled non-links and trains the model m_1 which EDoHa then uses to predict links between evidence and titles. After the user accepts one and rejects one link, they again train a new model. To keep the dataset balanced, EDoHa only needs to generate two weakly labelled non-links and train the model m_2 . The blue data represents the links and the red data the non-links.	103
7.5	Diagram of the ESIM used as EL model in EDoHa.	106
8.1	The five identified NLP tasks necessary to automatically validate hypotheses in a pipeline model. The tasks with the gray striped background are not considered in this thesis.	115
8.2	The document view of EDoHa with a medical abstract opened and manual annotations, as well as suggestions.	119
8.3	The evidence linking view of EDoHa with manual links and suggested links between claims and pieces of evidence. The numbers refer to the individual components described in the main text.	120

List of Tables

3.1	Overview of the existing datasets for ED.	24
3.2	Overview of existing datasets for CV.	29
4.1	The specific speeches used as the data source in our user studies.	41
4.2	Number of participants and users in both user studies.	42
4.3	The events logged in EDoHa to analyse user behaviour. The texts in quotes are constants.	46
4.4	The behaviour of the users with ED and EL phases and whether this user worked on exactly one hypothesis or document at-a-time.	51
4.5	Distribution of the users in our study according to their behaviour.	52
4.6	Number of revisions of the hypotheses for each user.	53
4.7	Responses of the users to the questionnaire.	54
4.8	The number of evidence annotations, hypotheses, and links each user created.	55
4.9	Agreement of the evidence of similar hypotheses (top), all hypothesis pairs with a substantial agreement on evidence (middle), and agreement of pre-defined hypotheses (bottom)	57
5.1	Hyper-parameters used for the direct training evaluation. The hyper-parameters are the number of hidden layers (HL), the size of the hidden layers (HLS), the learning rate (LR), the number of epochs (Ep), the batch size (BS), and the dropout (D).	64
5.2	Results of the nuclear energy dataset. The values for the evidence and no evidence rows are macro-averaged.	65
5.3	The values for each model are the evidence-F1 score for the nuclear energy dataset.	66
5.4	Results of the forest dieback dataset. The values for the evidence and no evidence rows are macro-averaged.	66
5.5	The values for each model are the evidence-F1 score for the forest dieback dataset.	66
5.6	Results of the forest dieback dataset after filtering the test documents that have been open for less than 20 seconds. The values in brackets are the difference with the results without the time filter.	68
5.7	Topic ids, topic titles, number of documents, and number of pieces of evidence in the ED-ACL-2014 dataset.	69
5.8	Statistics on the ED-EMNLP-2015 dataset.	70
5.9	Topic ids, topic titles, number of documents, and number of pieces of evidence in the ED-EMNLP-2015 dataset.	71
5.10	Statistics of the ED datasets after conversion to sentence-level annotated documents.	73
5.11	Model label depending on the training data.	74

5.12	Results of the pre-trained models on the test data of their source domain. The results are macro-averaged for F1, precision, and recall.	75
5.13	The results on the in-domain datasets are macro-averaged across all topics, with the standard deviations shown in parenthesis.	76
5.14	Number of documents, averages pieces of evidence per document, and evidence-F1 scores of the random baseline, $\text{bilstm}_{\text{direct}}$, $\text{bilstm}_{\text{fine}}$ models, and BERT for each topic on the ED-ACL-2014 dataset.	76
5.15	Number of documents, averages pieces of evidence per document, and evidence-F1 scores of the random baseline, $\text{bilstm}_{\text{direct}}$, $\text{bilstm}_{\text{fine}}$ models, and BERT for each topic of the ED-EMNLP-2015 dataset.	77
5.16	Evidence-F1 scores of the $\text{bilstm}_{\text{fine}}$ model and BERT for each topic of the UKP Sentential AM dataset.	78
5.17	Results on the 12 topics of the ED-EMNLP-2015, which are also present in the ED-ACL-2014 dataset. The scores are on the evidence class and macro-averages across the topics. The values in brackets are the difference from the ED-ACL-2014 dataset.	79
5.18	The results show only the evidence class and are macro-averaged across the three selected topics.	79
5.19	Number of documents and minimum number of training documents μ to reach a smaller error-rate E than BERT for the $\text{bilstm}_{\text{fine}}$ model for each topic on the ED-ACL-2014 dataset.	83
5.20	Number of documents and minimum number of training documents μ to reach a smaller error-rate E than BERT for the $\text{bilstm}_{\text{direct}}$ and $\text{bilstm}_{\text{fine}}$ models for each topic on the ED-EMNLP-2015 dataset.	84
6.1	Number of links and non-links per user of the nuclear energy (left) and forest dieback (right) datasets. For User12 in the nuclear energy and User10 in the forest dieback dataset, we could not create more unique non-links. . . .	89
6.2	Statistics of the ED-ACL-2018 dataset with the additional non-links. . . .	89
6.3	Group id, minimum and maximum number of links and non-links, and users for each group of both the nuclear energy and the forest dieback datasets. .	91
6.4	Statistics on the down-sampled datasets for the hyper-parameter tuning. . .	91
6.5	Possible values for the hyper-parameter optimisation.	92
6.6	Representatives of each user group for the nuclear energy and forest dieback datasets.	92
6.7	The different possible values for the number of epochs to train and the classification layer alone ilr , the number of epochs to train the entire network e , the learning rate in the initial training phase ilr , the learning rate when training the entire network lr , and the batch size.	92
6.8	Evaluation of the pre-trained EL models on the test data of the ED-ACL-2018 dataset. The results on the link and non-link classes are macro-averaged. .	93
6.9	The test results on the nuclear energy and forest dieback datasets.	94
6.10	User-specific results of our models in macro-F1 score on both classes on the nuclear energy dataset.	95
6.11	User-specific results of our models in macro-F1 score on both classes on the forest dieback dataset.	95
6.12	Evaluation of the upper bound of out-of-domain trained EL models for users in the nuclear energy and forest dieback data. The values are macro-F1 scores. .	96
7.1	Statistics on the datasets used to pre-train the ED and EL models.	104

7.2	Statistics on the datasets used to pre-train the argument detection and argument linking models.	104
7.3	Performance of the pre-trained models for evidence detection and evidence linking in English as well as German.	107
7.4	Performance of the pre-trained models for detection and linking on the AM data.	107
7.5	Answers of the student users participating the first user study.	110
7.6	Answers of the expert users to the quantitative questions of the questionnaire. The value 1 corresponds to the answer yes and the value 4 to the answer no.	111
C.1	Results of the individual hyper-parameter configurations of the bilstmAtt model. The labels for the different hyper-parameters are <i>lr</i> for the learning rate, <i>ep</i> for the number of epochs, and <i>bs</i> for the batch size.	146
C.2	Results of the individual hyper-parameter configurations of the bilstmAtt _{fine} model for the nuclear energy dataset. The different hyper-parameters are <i>lr</i> for the learning rate, <i>ilr</i> for the learning rate in the initial fine-tuning phase, <i>e</i> for the number of epochs, <i>ie</i> for the number of epochs in the initial fine-tuning phase, and <i>bs</i> for the batch size.	147
C.3	Results of the individual hyper-parameter configurations of the bilstmAtt _{fine} model for the forest dieback dataset. The different hyper-parameters are <i>lr</i> for the learning rate, <i>ilr</i> for the learning rate in the initial fine-tuning phase, <i>e</i> for the number of epochs, <i>ie</i> for the number of epochs in the initial fine-tuning phase, and <i>bs</i> for the batch size.	148

Anmerkungen zum Umgang mit Forschungsdaten

Gemäß der “Leitlinien zum Umgang mit Forschungsdaten” der Deutschen Forschungsgemeinschaft¹ wurden alle im Zusammenhang mit dieser Dissertation entstandenen Forschungsdaten langfristig archiviert und sofern möglich öffentlich zugänglich gemacht. Folgende Forschungsdaten wurden frei verfügbar gemacht:

- **Software & Korpora**

- EDoHa ist unter der Apache License Version 2.0 veröffentlicht und kann unter der URL <https://github.com/UKPLab/EDoHa> im Quelltext heruntergeladen werden. Dies enthält auch den Quelltext mit dem die vortrainierten Modelle aus Kapitel Interactive Evidence Detection erstellt wurden.
- Der Quelltext der Experimente aus dem Kapitel Machine Learning for Evidence Detection befinden sich in dem Repository <https://github.com/UKPLab/fever2019-interactive-evidence-detection> unter der Apache Licence Version 2.0. Dort befinden sich auch die Korpora in der Version in der sie in den Experimenten genutzt wurden. Sofern die Verbreitung der Korpora aus lizenzrechtlichen Gründen nicht möglich ist, liegt der für die Erstellung der Korpora notwendige Quelltext in dem Repository.

- **Forschungsergebnisse**

- Alle im Zusammenhang mit dieser Dissertation stehenden Publikationen sind in der Präambel dieser Dissertation verlinkt.
- Alle Forschungsergebnisse sind zudem auch in dieser Dissertation selbst dokumentiert, die von der Universitäts- und Landesbibliothek Darmstadt zur Verfügung gestellt wird.

Weitere in dieser Dissertation beschriebene Datensätze können aus urheberrechtlichen und datenschutzrechtlichen Gründen nicht frei verfügbar gemacht werden. Entsprechend der DFG-Leitlinien sind diese Daten sowie damit zusammenhängende Software intern unter Nutzung der Infrastruktur der Universitäts- und Landesbibliothek Darmstadt archiviert, so dass eine Archivierung für mindestens 10 Jahre gewährleistet ist. Die vortrainierten Modelle können auf Anfrage herausgegeben werden.

¹https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

Ehrenwörtliche Erklärung ²

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades “Dr. rer. nat.” mit dem Titel “Interactive Evidence Detection” selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 05.06.2020

Chris Stahlhut, M.Sc.

²Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt